OPEN Open Patient data Explorative Network

GUIDE TO REDCAP EXPORTED FILES

UNDERSTANDING DATA FORMATS AND LOADING DATA INTO ANALYSIS SOFTWARE

INDHOLD

INTRODUCTION

At some point in time in the course of your REDCap project, you will need to export your data from REDCap, in order to do your statistical analysis. This may be at the end of your data collection phase, or it may already be during data collection. REDCap is capable of letting you design your exports in various ways (as described in a separate guide), but it is not always easy to know how to manage and interpret the actual files exported from REDCap. This guide will hopefully be able to help you in this phase of your project

REDCAP DATA FILES

As covered in the guide to the REDCap export module, REDCap allows the user to export data in specific formats for a number of different analysis software (Excel, Stata, SAS, SPSS and R). For each of these, the download consists of at least 2 different files.

For Stata, SAS, SPSS and R the set of files contains 1 data file, 1 script file and for SAS and SPSS an additional pathway mapper file. The data files are normal text files which could essentially be opened in notepad (not recommended). They contain the data as one row pr recordid/event combination and with the data values for each variable (column) next to each other separated by commas (the .csv suffix of the files stands for comma-separate-values). The csv files are sensitive, as they contain the actual data.

The script files are sets of instructions for SAS, R, STATA or SPSS, written in the syntax of those softwares. In STATA this is known as a Do-file. The instructions load the data file, and adds all of the metadata for each of the variables/columns in the project. That is, it tells e.g. STATA what the field label for a given variable is, and which coded options (for e.g. a radiobutton field) were available and the labels for those options. The pathway mapper files for SAS and SPSS are small programs that should be executed first, in order for the script file to know the name and location of the data file.

Excel is a bit different from the other software packages, in that both of the files available for download are data files. There is no script file. It is not possible in Excel to associate metadata to each variable. Instead the two data files represent data in two different ways. The one called 'raw' contains the data in a way where all column names are named by the variable name and where data points are shown with the code for the given choice (e.g. 1 for yes). In the file called 'label' columns are named with the field label and data points are shown with the label for the choice (e.g. 'yes')

STORING REDCAP DATAFILES

The recommended procedure for downloading data from REDCap, is to download all the export files from the selected software package (e.g. STATA) into the same folder. This folder should be placed in a secure environment with logging and access control. In the OUH/RSD setting, we recommend that this is a secure SharePoint folder. OPEN has written a dedicated guide to setting up a SharePoint site and mapping folders from this site on your local pc.

A mapped SharePoint folder will seem like a normal folder on your own machine. As you work with your data, make sure you only open and save from this folder. In that way, your access to the data files will be logged and your analysis phase will comply with the regulations defined by the data protection agency. Note however, that this is only true if the data files you work on in your analysis phase are de-identified, i.e. do not contain identifiers such as CPR-numbers, names, etc.). For this reason we recommend that downloads/exports from REDCap should generally not contain variables with identifiers. Typically these are not parameters you will need in your analysis anyway. Keep those inside REDCap where they are safe, available for look-up and logged according to the special regulations that exist for identifiable data. For more information about SharePoint and identifiable data, see OPENs SharePoint guide.

If during the download process you temporarily store files somewhere on your local machine (e.g. in a download folder), make sure to clean up by deleting those files after you move them to the secure location. In windows you need to press shift+delete in order to delete completely and not add to trash bin.

UNDERSTANDING THE REDCAP DATA FORMAT

There are two different ways REDCap formats the output data depending on whether the project is set up as longitudinal or traditional.

TRADITIONAL PROJECT

In a traditional project, every single field created in the project has its own variable name, and data is only entered for that variable one time for each record ID in the course of the project. In this setup it is possible to represent all the exported data points for a given individual in a single row, with one column for each variable (note special case for checkbox variables covered later in this guide). Table 1 below illustrates this point in an example project containing 14 variables on three instruments (baseline data, visit 1 and visit 2). Each patient (RecID) has all his/her data represented in a single row. Another way to say this, is that a given data point/value can be uniquely pinpointed with just 2 pieces of information (coordinates): The variable name (which data element), the record ID (who does the data element belong to).

Table 1

	Baseline data						Visit 1				Visit 2			
RecID	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14
pt_01	value	value	value	value	value	value	value	value	value	value	value	value	value	value
pt_02	value	value	value	value	value	value	value	value	value	value	value	value	value	value
pt_03	value	value	value	value	value	value	value	value	value	value	value	value	value	value
pt_04	value	value	value	value	value	value	value	value	value	value	value	value	value	value

LONGITUDINAL

When the project is set up explicitly longitudinal in REDCap, each field and variable is potentially involved at multiple data entry events, where the same instrument is reused. REDCap does of course not overwrite the existing value for the variable, when it is entered at a later event. Instead it uses the event name as an extra piece of information to pinpoint a data point/value uniquely. So in addition to the variable name and record id mentioned above, we also need to know the event name to know which instance of the variables use, we refer to. This is added as an extra column to the exported data called 'redcap_event_name', and the different events are represented on separate rows for each individual. In the table below this is illustrated for the same project as before, except that now the project is set up longitudinally. The project has three defined events: a baseline event where only the baseline instrument is filled and two visit events where only the visit instrument is filled.

Table 2

		Baseline data						Visit			
RecId	redcap_event_name	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10
pt_01	baseline_arm_1	value	value	value	value	value	value				
pt_01	visit_1_arm_1							value	value	value	value
pt_01	visit_2_arm_1							value	value	value	value
pt_02	baseline_arm_1	value	value	value	value	value	value				
pt_02	visit_1_arm_1							value	value	value	value
pt_02	visit_2_arm_1							value	value	value	value

We see that each patient (RecId) is now represented in multiple rows. The 4 variables of the visit instrument are filled 2 times each, and that we can keep track of which instance each value belong to by correlating it with event name in the 'redcap_event_name' column. We also see how all the variables in the baseline instrument are only filled one time, and that we have empty values for the events where they were not used. The same goes for the visit variables in the baseline event. This is what the exported data looks like in a longitudinal project and it can perhaps seem a bit confusing at first.

Depending on the status of your data collection, when you perform your export, you may experience that your patients do not have the same number of rows in the exported data. This is because some of the patients may not have had any data saved for some of the events. In the example in Table 2 both patients have three rows of data, but if e.g. pt_02 had not yet been seen for visit 2, then the row with the event name visit_2_arm_1 would not be part of the export.

INSTRUMENT STATUS VARIABLE

There are some special variables in REDCap, that are not often used within REDCap, but that are part of the exported dataset. For each instrument in the project, REDCap creates a variable that gets the name of the instrument (with underscore for spaces) followed by the word complete. So for an instrument called 'baseline', REDCap create a variable called baseline_complete. This variable is used in REDCap to track the save-status for a given instrument-event-patient combination, corresponding to the color codes used in REDCap. The variable can take the values 0,1,2 corresponding to 'incomplete (red)', unverified (yellow) and complete (green).

CHECKBOX FIELDS

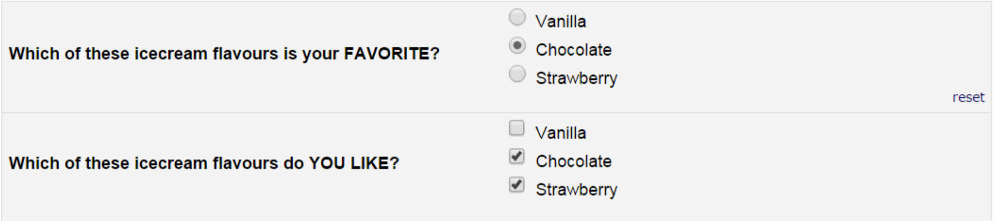
Even though checkbox fields seem similar to dropdown fields and radiobutton fields in REDCap, they are stored and managed in a different way in the database, and have a different representation in the exported data.

A radiobutton field (or dropdown) accepts only a single answer, and since each possible answer is associated to a code, the data point of a filled radiobutton field can be represented as a single value in a single column/variable in the exported data.

A checkbox field can take as many answers as there are choices in the checkbox field. It is not obvious how the many codes that these answers represent, should be stored in a single column, and still make sense analytically. What REDCap does (and this is the standard way to do this), is to represent each possible answer as a separate binary variable/column and indicate 1 or 0 (checked or unchecked) as the value. This means, that if your project contain checkbox fields, your exported data will contain more columns/variables than you originally designed.

The additional columns/variables that this produces are named with the variable name given to the checkbox field in the design phase and then a suffix consisting of two underscores followed by the code for each choice, e.g. variable_name__1.

This is illustrated in **Fejl! Henvisningskilde ikke fundet.** and Table 3 below. In both the radiobutton field and the checkbox field, the choices are coded as: 1=Vanilla, 2=Chocolate, 3=Strawberry.



Which of these icecream flavours is your **FAVORITE**?
○ Vanilla
● Chocolate
○ Strawberry

Which of these icecream flavours do **YOU LIKE**?
□ Vanilla
☑ Chocolate
☑ Strawberry

Table 3

RecordID	Icecream_favorite	Icecream_like__1	Icecream_like__2	Icecream_like__3
1001	2	0 (unchecked)	1 (checked)	1 (checked)

Figure 1

MISSING VALUES

You may have designed your project in a way that have allowed you to explicitly code missing data values, rather than just leaving them blank (e.g. having a choice for 'missing' in a radiobutton field). But quite often you will also have fields that you simply did not fill out during data entry for various reasons. When REDCap creates exports, these non-filled fields are represented in different ways. For most field types (text boxes, notes boxes,..), a missing value will be represented as missing. In Excel that means it will be represented as a blank and in STATA as a dot '.'

RADIOBUTTONS AND CHECKBOXES

Importantly radiobuttons and checkboxes have different ways of representing a missing value in exported data. A radiobutton field where none of the buttons have been clicked will have its value exported as 'missing' (empty field in Excel, "." In Stata) in the same way as a textbox field. A checkbox field however will create as many new binary variables as there are choices, and each of these new variables will contain either a 0 (un-checked) or a 1 (checked), even if all the boxes are unchecked. So an unchecked checkbox is 0 and an unchecked radiobutton is missing.

SLIDERS

A slider field that has not been touched at all, will be represented as missing, but if the slider has been touched (even when returned to the starting position), it will store the corresponding value.

FIELDS HIDDEN BY BRANCHING LOGIC

Another way, in which a data value can be missing in REDCap, is if that variable/field was never exposed to the entry person because of branching logic (e.g. a question about pregnancy being hidden for men). In systems without branching logic or paper based data entry, such fields would typically be coded explicitly as 'not relevant'. In REDCap the entry person is spared having to do this because of the branching. Unfortunately REDCap does not have any special way to represent this in the exported data, so a field that was hidden by branching for a particular patient, and therefore not filled, will simply be represented as missing in the exported data (blank in Excel, '.' in STATA). This is something you may have to account for in your analysis, so you don't end up counting 'not relevant' as missing.

LOADING DATAFILES

Above we described the structure of the exported data. Next we demonstrate how you may import you data files into you analysis software. The procedure for opening downloaded data files differs depending on which software is used. This guide will only describe the procedure for Excel, OpenOffice and Stata.

EXCEL

As already described, the exported data files from REDCap are .csv files, and this is also the case for those that REDCap offers as Excel files. They are not truly Excel files (that end in .xls). In order to make sure that Excel understands how to interpret the datafile, it is recommended to open the file using data import rather than the normal 'open file' dialogue. This is shown below for Excel 2010 (dansk).

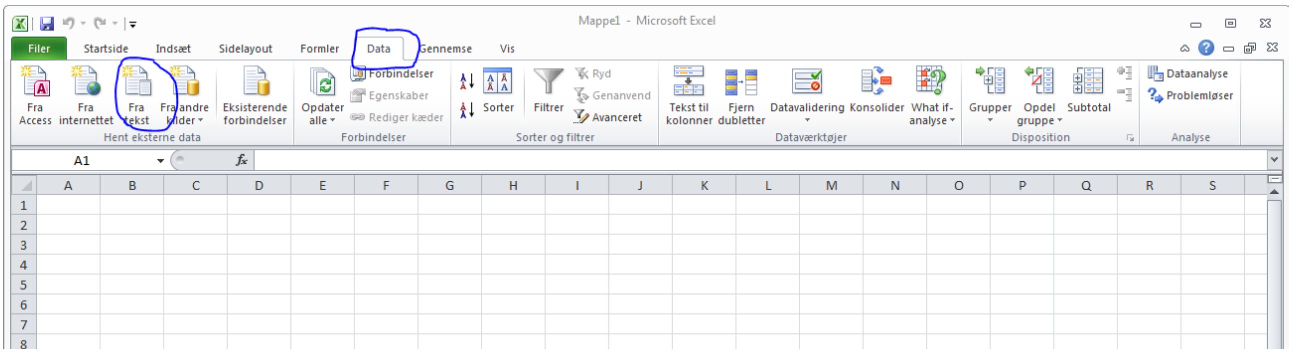
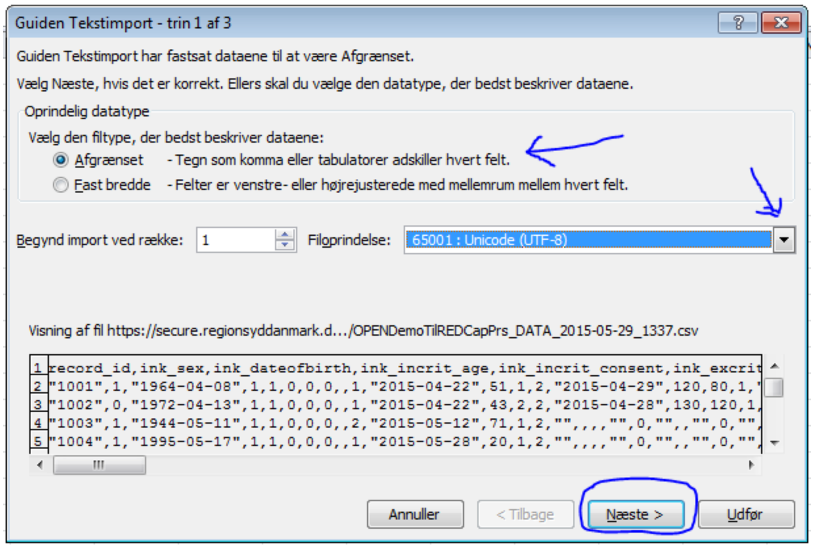


Figure 2

From the 'Data' tab choose 'Fra tekst' (Figure 2). In the next window find the .csv file you wish to open and press 'Åbn'. If you have a mapped SharePoint folder where you store your pseudo anonymized analysis files, this is the folder you should navigate to in order to find your file. In the next dialogue (Figure 3) we tell Excel, that the columns in the dataset are delimited by a certain character (rather than a fixed width). We also choose the file encoding to be Unicode (UTF-8) to ensure proper representation of special characters like 'æøå' and press 'næste'


Guiden Tekstimport - trin 1 af 3

Guiden Tekstimport har fastsat dataene til at være Afgrænset.

Vælg Næste, hvis det er korrekt. Ellers skal du vælge den datatype, der bedst beskriver dataene.

Oprindelig datatype

Vælg den filtype, der bedst beskriver dataene:

● Afgrænset - Tegn som komma eller tabulatorer adskiller hvert felt.

○ Fast bredde - Felter er venstre- eller højrejusterede med mellemrum mellem hvert felt.

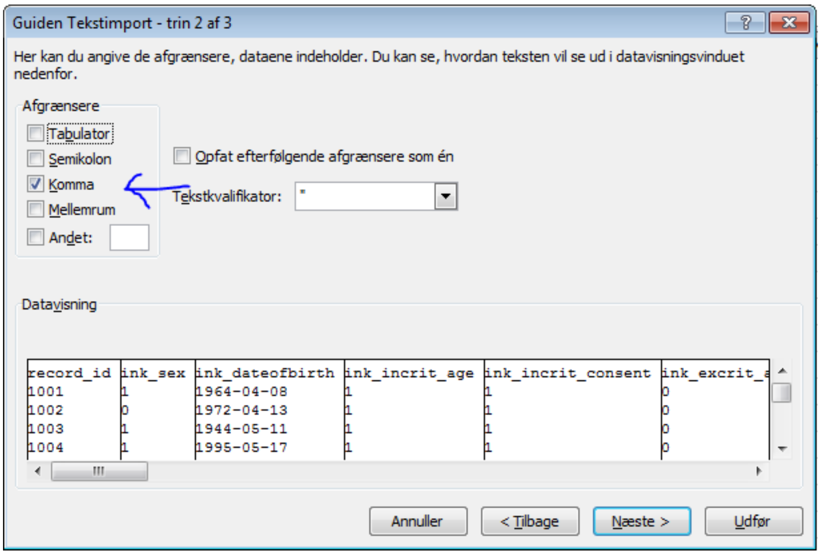
Begynd import ved række: 1 Filoprindelse: 65001 : Unicode (UTF-8)

Visning af fil https://secure.regionsyddanmark.d.../OPENDemoTilREDCapPrs_DATA_2015-05-29_1337.csv

```
1 record_id,ink_sex,ink_dateofbirth,ink_incrit_age,ink_incrit_consent,ink_excrit
2 "1001",1,"1964-04-08",1,1,0,0,0,,1,"2015-04-22",51,1,2,"2015-04-29",120,80,1,"
3 "1002",0,"1972-04-13",1,1,0,0,0,,1,"2015-04-22",43,2,2,"2015-04-28",130,120,1,
4 "1003",1,"1944-05-11",1,1,0,0,0,,2,"2015-05-12",71,1,2,"",,,,"",0,"",,"",0,"",
5 "1004",1,"1995-05-17",1,1,0,0,0,,1,"2015-05-28",20,1,2,"",,,,"",0,"",,"",0,"",
```

Figure 3

In the next step (Figure 4) we specify that the columns are delimited by commas and press 'næste'.



Guiden Tekstimport - trin 2 af 3

Her kan du angive de afgrænsere, dataene indeholder. Du kan se, hvordan teksten vil se ud i datavisningsvinduet nedenfor.

Afgrænsere

☐ Tabulator
☐ Semikolon
☑ Komma
☐ Mellemrum
☐ Andet:

☐ Opfat efterfølgende afgrænsere som én

Tekstkvalifikator: "

Datavisning

```
record_id ink_sex ink_dateofbirth ink_incrit_age ink_incrit_consent ink_excrit_
1001      1       1964-04-08      1              1                  0
1002      0       1972-04-13      1              1                  0
1003      1       1944-05-11      1              1                  0
1004      1       1995-05-17      1              1                  0
```

Figure 4

In the final step it is possible to specify the datatypes of each of the datacolumns or even to skip certain columns. This can also be done later after the data has been loaded or it can be skipped. Press ‘udfør’ and choose the upper left to import the data into.

ISSUES WITH EXCEL

If the imported data looks like you expect, then all is well. However, in quite a few projects this is not the case, and columns will look strange. This is most often the case with the ‘label’ version of the exported files. The problem is that Excel tends to confuse the line shifts (‘enter’) that may have been used in long field labels or elsewhere, with new records, and therefore breaks the rows in the wrong places. There is unfortunately no easy fix for this, and rather than spending a lot of time manually rearranging your data, we recommend that you instead use the spreadsheet program called ‘Calc’ from the OpenOffice suite of office programs. OpenOffice is similar to Microsoft Office, but is open source and can hence be downloaded freely from

https://www.openoffice.org/

Another similar alternative is libreoffice. Both of these will run on both Windows, Mac and Unix machines, and have no problems interpreting the REDCap data files. They will also be the preferred choice for working with the REDCap datadictionary, if you use this as a supplement to the online designer tool in the design phase of your REDCap project

IMPORTING DATA USING OPENOFFICE CALC

Assuming you have already downloaded and installed OpenOffice, open up OpenOffice and choose the Spreadsheet option ‘Calc’ (Figure 5).

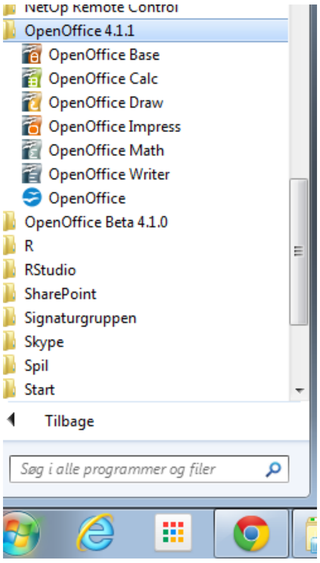
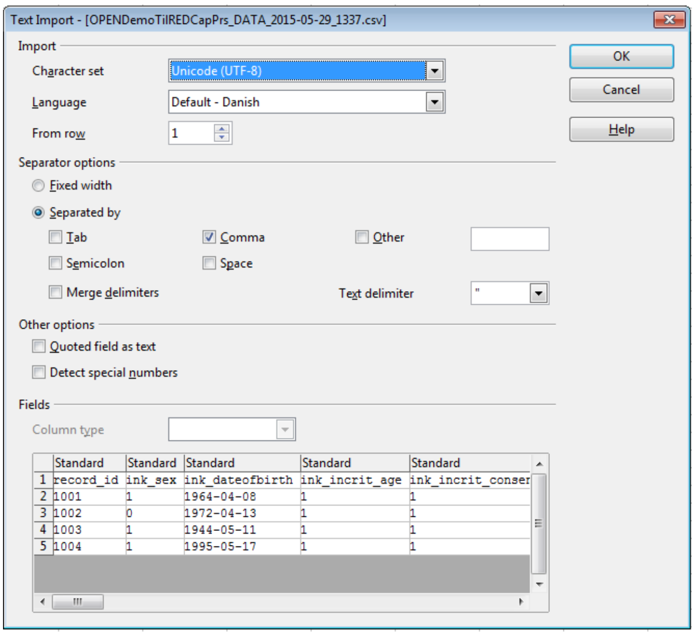


Figure 5

From the top left corner press file and choose 'Open'. Navigate to the file you wish to open and press 'open/åbn'. OpenOffice will see that you are trying to import a .csv file and open the 'Text Import' dialogue (Figure 6). Most often it will correctly guess that columns are separated by commas and that the character set is Unicode (UTF-8). If not, set the values as in the figure below (Figure 6) and press ok.


Text Import - [OPENDemoTilREDCapPrs_DATA_2015-05-29_1337.csv]

Import

Character set: Unicode (UTF-8)

Language: Default - Danish

From row: 1

Separator options

○ Fixed width

● Separated by

□ Tab ☑ Comma □ Other

□ Semicolon □ Space

□ Merge delimiters Text delimiter: "

Other options

□ Quoted field as text

□ Detect special numbers

Fields

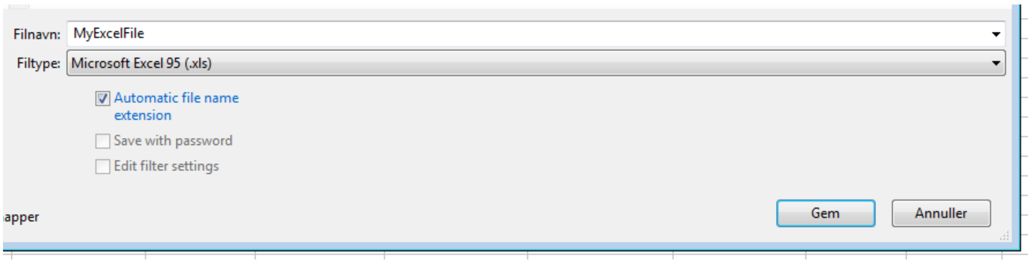
Column type

	Standard	Standard	Standard	Standard	Standard
1	record_id	ink_sex	ink_dateofbirth	ink_incrit_age	ink_incrit_conser
2	1001	1	1964-04-08	1	1
3	1002	0	1972-04-13	1	1
4	1003	1	1944-05-11	1	1
5	1004	1	1995-05-17	1	1

Figure 6

Your data should be imported and look the right way. You can now choose to stay in OpenOffice and do your analysis, or you can save the data you just imported in Excel format, and then open the file in Excel as a normal Excel file, if you prefer that. OpenOffice will keep the file in .csv format unless you choose to save it as a .odf file, which is the

OpenOffice spreadsheat file format. To save in Excel format click ‘file’ in upper left corner and press ‘save as’. Give the file a name and choose the file type to be Microsoft Excel 95 (.xls)(Figure 7).



Filnavn: MyExcelFile

Filtype: Microsoft Excel 95 (.xls)

☑ Automatic file name extension

□ Save with password

□ Edit filter settings

Figure 7

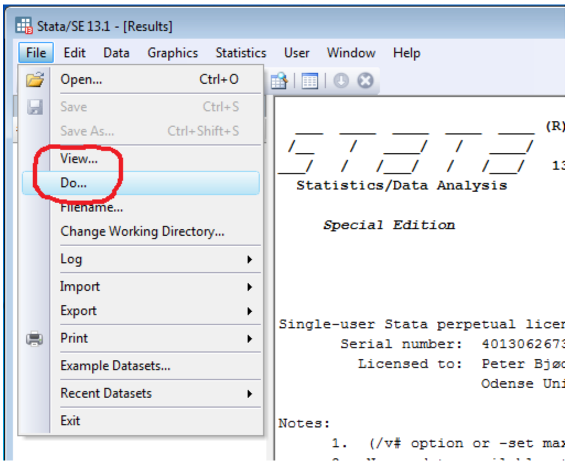
This file can now be opened in Excel

IMPORTING DATA USING STATA

If you have chosen STATA as the export format from REDCap, you should have two downloaded files. A data file with a .csv suffix, and a STATA do-file. Place these two files in the same folder (e.g. a mapped SharePoint folder). You now need to run the STATA do-file in order to import the data file and annotate it with all the information about your variables.

RUNNING A DO-FILE

This can be done in a number of different ways, but one way is to open STATA and under file in the upper left corner press ‘do...’ (Figure 8)

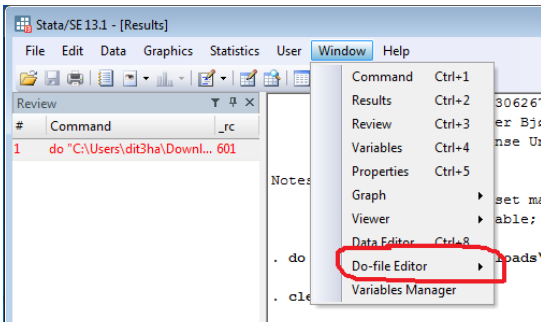


```
  ___  ____  ____  ____  ____  (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   13
  Statistics/Data Analysis

    Special Edition

Single-user Stata perpetual licer
       Serial number:  401306267
         Licensed to:  Peter Bjø
                       Odense Uni

Notes:
      1.  (/v# option or -set ma
```

Figure 8

In the dialogue that opens navigate to the do-file you downloaded from REDCap. This should execute the commands in the do-file which loads your data file and annotates it.

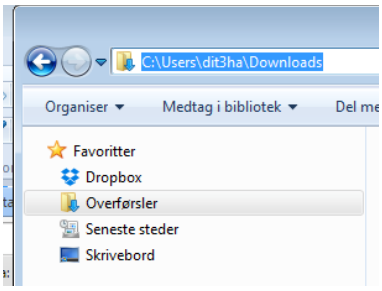
You may encounter an error message that says: file filename.csv not found. This happens if the do-file is incorrectly pointing to the wrong path/address of the data file. If this happens you should open up the do-file in STATAs do-file editor (Figure 9)



1 do "C:\Users\dit3ha\Downl... 601

Figure 9

In the editor window that opens, open the do-file by choosing file -> open. This will open up the do-file and show you all the commands it contains. The second line in the file begins with the word ‘insheet’, and it is here that the do-file is told where to look for the data file. If you scroll to the right end of that line (this line may be very long), there is a final statement that says: ‘using “Filename.csv”, nonames’. We need to add the absolute path to the .csv file in front of the filename. You can find the absolute path of the .csv file, by opening up your windows explorer (stifinder) and navigate to the location of the file. When you find the file, click in the address field at the top (Figure 10) and copy the path in front of the data file to give STATA the full address of the file. The resulting statement in the do-file should be something like ‘using C:\Users\username\Documents\Filename.csv, nonames’. Once this is changed, you should be able to execute the do-file.



C:\Users\dit3ha\Downloads

Figure 10

Once you have your dataset loaded and properly annotated in STATA, you can save it as a standard STATA file for your continued analysis. Remember to always open and save your data files from and to a secure location, preferably a mapped SharePoint folder.

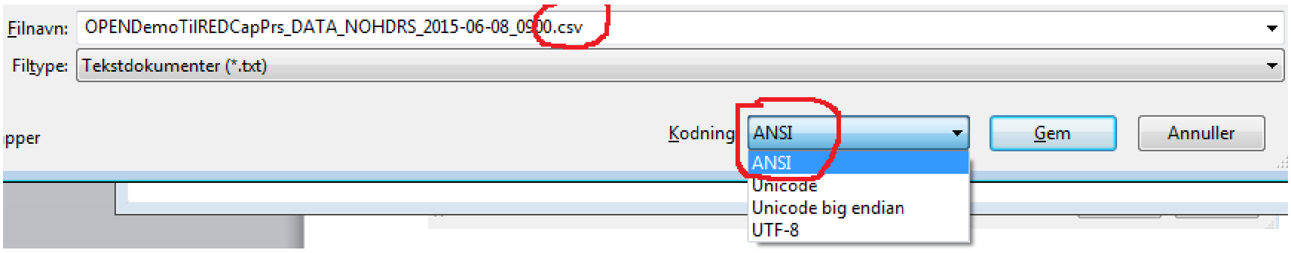
SPECIAL CHARACTERS IN STATA VERSION 13 AND EARLIER

If you have used special characters like æøå in your REDCap project, e.g. in your field labels, or because free text answers have used them, you will experience that these characters are not represented correctly in versions of STATA earlier than 14. This means that e.g. æøå will be represented by a strange combination of characters. This is an issue with STATAs understanding of the UTF-8 file encoding that REDCap uses. This can be a bit annoying, but should not be a major issue for you analysis. It can be solved by first opening the data file in a program that is able to correctly read the encoding of the data file and do-file, and then saving them again in another encoding that STATA understands. This is not something you have to do if you don't have any problems with the representation.

RE-ENCODING STATA FILES USING NOTEPAD

The simplest tool to use for re-encoding is the Windows notepad (notesblok) program. On Mac you will likely have a similar program that can do this. Use this procedure for re-encoding the data file and do-file.

1. Open the notepad program
2. Select Open file from the 'Filer' menu and navigate to the data file or do-file that you need to re-encode.
    a. You will probably need to select 'alle filer' to get it to list .do files and .csv files
    b. Check that the dropdown menu called 'Kodning' is set to UTF-8
3. The file should now open. If this is the data file it will probably look a bit confusing, but this is not important. Don't change anything in the file
4. Select 'save as' (gem som) from the 'Filer' menu (Figure 11)
    a. In the 'save as' dialogue there is a dropdown menu called 'Kodning'. Choose ANSI
    b. You have to manually add the right ending to the file. .csv for your data file and .do for the do-file
    c. Don't change the name of the data file, since the do-file expects the existing name



Filnavn: OPENDemoTilREDCapPrs_DATA_NOHDRS_2015-06-08_0900.csv

Filtype: Tekstdokumenter (*.txt)

Kodning: ANSI

Figure 11

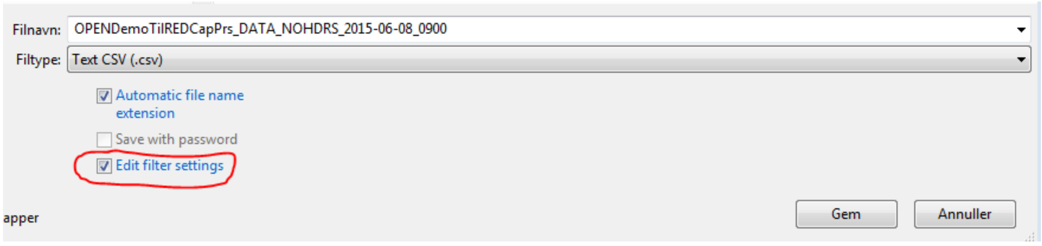
5. Save the file overwriting the existing file
6. Follow the same procedure for the other file

The resulting files should now be encoded in the right way for STATA to properly load them (using the procedure described earlier) and understand the special characters.

RE-ENCODING STATA FILES USING OPENOFFICE CALC

OpenOffice Calc is another program that can do this. If you have a very large data file you may experience problems with the notepad method. Also, Calc allows you a more detailed choice of encodings and is available on Mac and Linux.

1. Open the data file using the same procedure as already described for loading a csv file into Calc.
2. Once it is open, click 'save as' from the file menu. This open the save as dialogue (**Fejl! Henvisningskilde ikke fundet.**)



Filnavn: OPENDemoTilREDCapPrs_DATA_NOHDRS_2015-06-08_0900

Filtype: Text CSV (.csv)

☑ Automatic file name extension

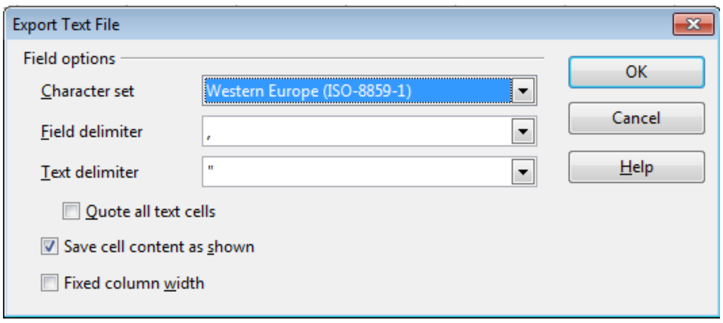
□ Save with password

☑ Edit filter settings

Figure 12

a. Keep the same name and save to the same location
b. Remember to click the 'edit filter settings' checkbox (**Fejl! Henvisningskilde ikke fundet.**). This will allow you to change encoding in a later dialogue.

3. Say yes to overwrite the existing file
4. Choose 'keep current format' in the next dialogue
5. You are now given options about how to encode the file and how to delimit the columns (**Fejl! Henvisningskilde ikke fundet.**). In the Character set drop-down, choose the option 'Western Europe (ISO-8859-1)'. There are also other options that will work, but the ISO-8859-1 is a standard you should be able to find in other software and on other operating systems like Mac and Linux. Do not change any of the other values.



Export Text File

Field options

Character set: Western Europe (ISO-8859-1)

Field delimiter: ,

Text delimiter: "

□ Quote all text cells

☑ Save cell content as shown

□ Fixed column width

Figure 13

6. Press ok. The new data file should now represent special characters correctly when loaded into STATA