

Researcher Handbook

IPUMS International Research Data Enclave

Updated: 2018-06-26

Contents

- Preface** **2**

- 1 Applying for access** **3**
 - 1.1 Eligibility 3
 - 1.2 Forms to complete 3

- 2 Analyzing the data** **4**
 - 2.1 Logging in to the remote desktop 4
 - 2.2 Logging off the remote desktop 5
 - 2.3 Navigating the file system 6
 - 2.4 Loading the data into statistical software 9
 - 2.5 Using Stata and R packages 11

- 3 Reviewing output for disclosure risk** **12**
 - 3.1 Disclosure protection rules 12
 - 3.2 Preparing output for review 13
 - 3.3 Modifying disclosive output 14
 - 3.4 Requesting review of output for release 14
 - 3.5 Output release procedure 15

- 4 Concluding a project** **16**
 - 4.1 Archiving project files 16
 - 4.2 Citing the IPUMS IRDE 16

Preface

This guide is for users of the IPUMS International Research Data Enclave (IRDE). It is meant to guide you through completing an application for access to data hosted by the IRDE, and using the IRDE's remote desktop environment to complete your analyses. As a user of the IPUMS IRDE, you are responsible for knowing the rules and policies explained in this handbook.

The IRDE was established to enable research that is only possible using confidential microdata that cannot be made widely available. To prevent disclosure of confidential information, IRDE data must be accessed through a secure web portal, and no analysis output can be removed from the IRDE virtual environment without being reviewed by IRDE staff. Moreover, to be granted access to IRDE data, you the researcher must pledge to protect the confidentiality of the data, and take all appropriate steps to prevent disclosure.

This guide is organized as follows.

Chapter 1 describes what types of researchers and research projects are eligible for access to IPUMS IRDE data, and walks through the application process.

Chapter 2 provides instructions for accessing and navigating the virtual environment in which researchers can analyze IRDE data.

Chapter 3 focuses on the rules for what types of output can be released from the virtual environment, and describes the procedures for preparing output and submitting a request for IRDE staff to review your output.

Chapter 4 describes procedures for concluding your IRDE project, including how we archive your analysis files and how you should cite the IPUMS IRDE.

Chapter 1

Applying for access

1.1 Eligibility

To be eligible for access to data hosted on the IRDE, a research project generally must be led or supervised by a faculty member at an accredited college or university, or by a researcher at a not-for-profit, non-governmental organization who possesses a doctoral degree. Independent researchers not affiliated with a research institution, or researchers working for private, for-profit firms are not eligible for access to IRDE data. If you are unsure of your eligibility, please contact the IRDE staff at ipums-irde@umn.edu.

In addition to these requirements related to institutional affiliation, you must also demonstrate that your research questions *can* be addressed using data hosted on the IRDE, and that they *cannot* be addressed with data available elsewhere. In particular, you must be able to explain clearly why your research questions cannot be answered using non-restricted IPUMS International data.

Finally, to be eligible for access to the IRDE, a project must not pose substantial risk of disclosure of confidential information. Identifying information about low-level geographies or other small population subgroups will be removed from analysis results during the output review process. If answering your research question would require disclosure of such information, your project will not be eligible for access. Again, please contact the IRDE staff (ipums-irde@umn.edu) with any questions.

1.2 Forms to complete

To apply for access, you must complete the *Application to Use Restricted Data*. The *Application to Use Restricted Data* consists of two parts, one for providing personal information about the researcher(s) applying for access, and one for providing information about the research project.

Each member of the research team who will be accessing the IRDE virtual environment must fill out a copy of the personal information form, and the principal investigator (PI) must additionally complete the project information portion of the *Application to Use Restricted Data*.

If a project is approved for access to the IRDE, each member of the research team must read and sign the *Confidentiality Pledge*, and both the PI and a representative of the PI's institution must sign the *Confidentiality Agreement*.

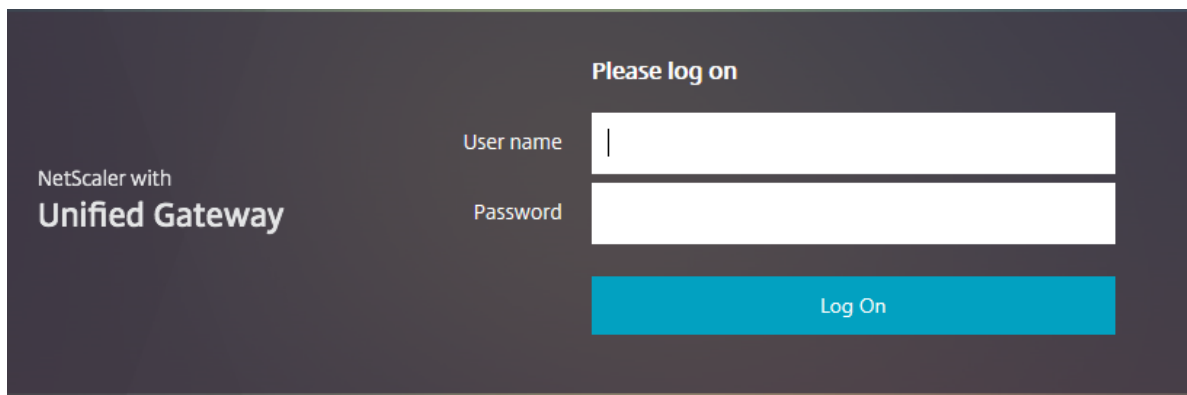
Chapter 2

Analyzing the data

2.1 Logging in to the remote desktop

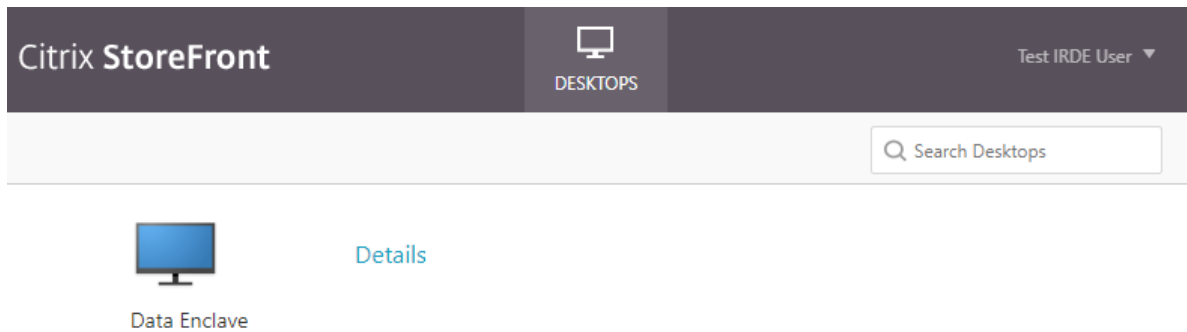
You are only able to log in to the IRDE virtual environment from designated computers at IRDE partner centers. Once you have been granted access to a designated computer at an IRDE partner center, follow these instructions on the computer to access the virtual environment.

1. Navigate to <https://irde.pop.umn.edu>.
2. Enter your login credentials.

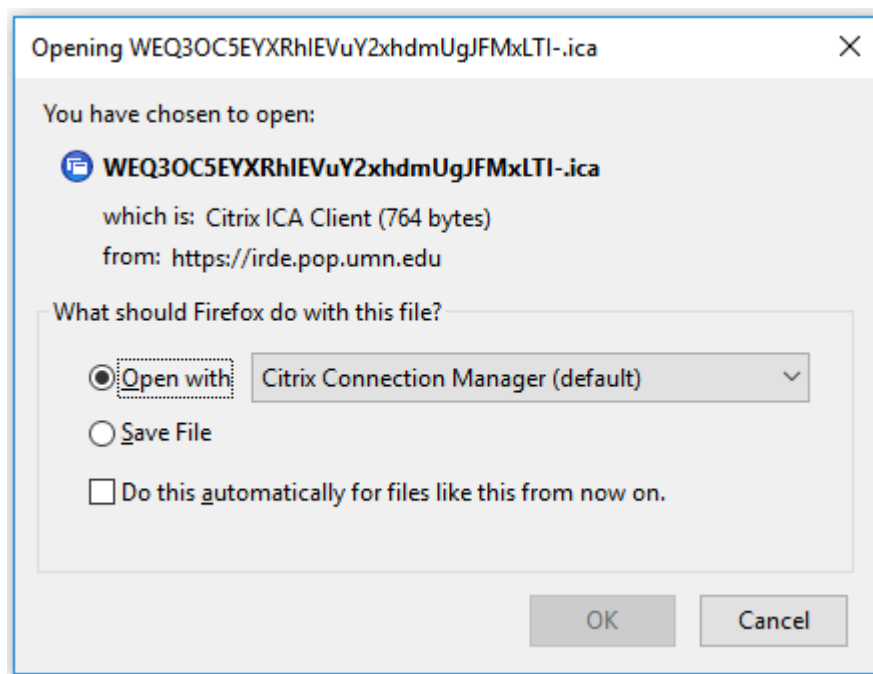


The image shows a login interface for NetScaler with Unified Gateway. The background is dark gray. On the left, the text "NetScaler with Unified Gateway" is displayed in white. On the right, the text "Please log on" is displayed in white. Below this text, there are two white input fields: one for "User name" and one for "Password". A blue button labeled "Log On" is positioned below the password field.

3. Click on the monitor icon labeled "Data Enclave".



4. Make sure that the “Open with” option is selected, and that “Citrix Connection Manager (default)” is selected in the program dropdown, then click “OK” to launch the remote desktop environment.

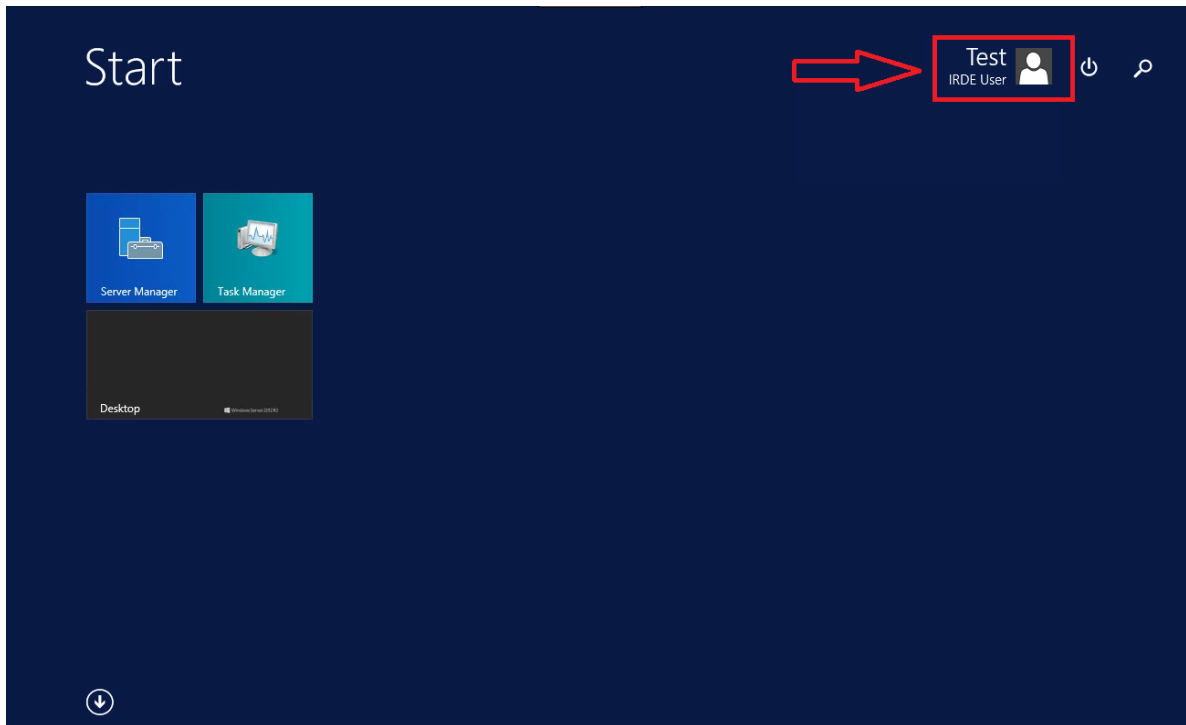


2.2 Logging off the remote desktop

When you are finished working in the remote desktop environment, always follow these steps to log off. **Failure to follow these steps may lock the remote desktop and prevent you or other users from logging back on.**

1. Click the Windows icon in the bottom left corner of the screen to open the Start menu.

2. Click the user icon in the top right corner of the start menu, as highlighted in the screenshot below:

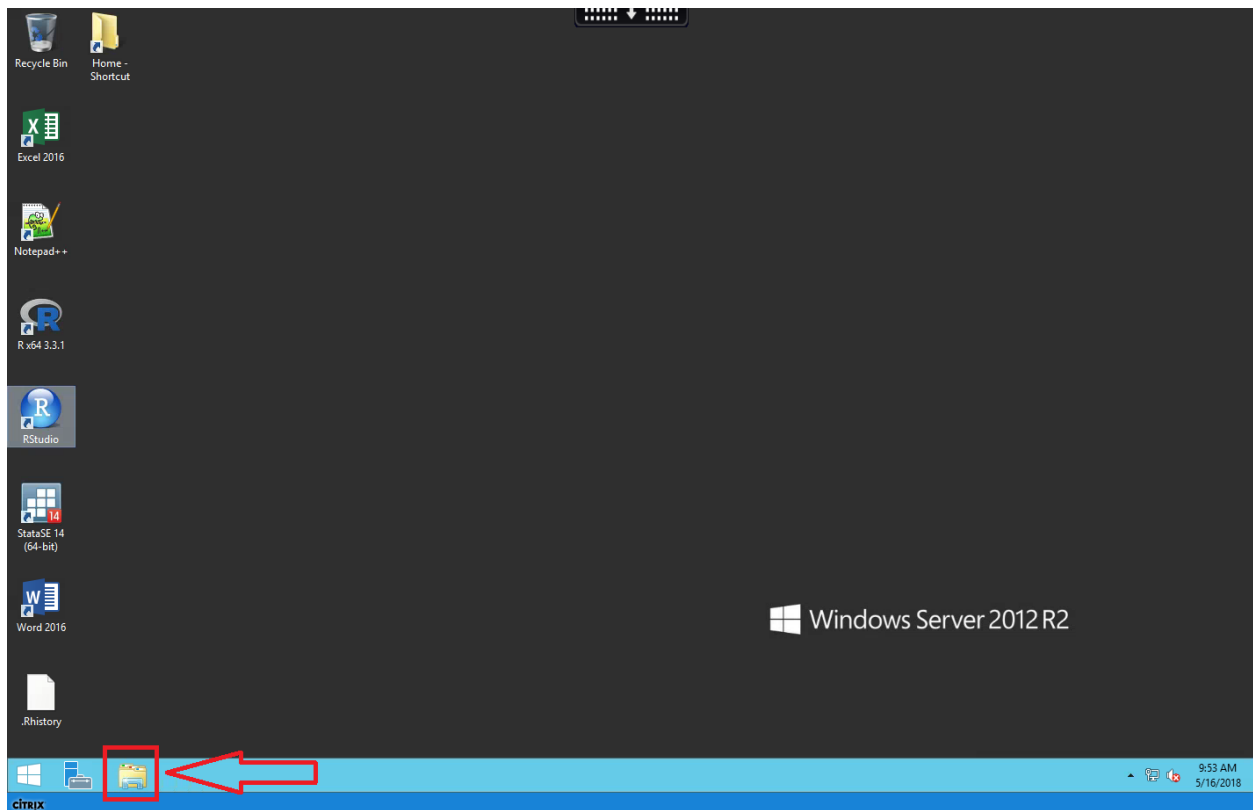


3. Click "Sign out".

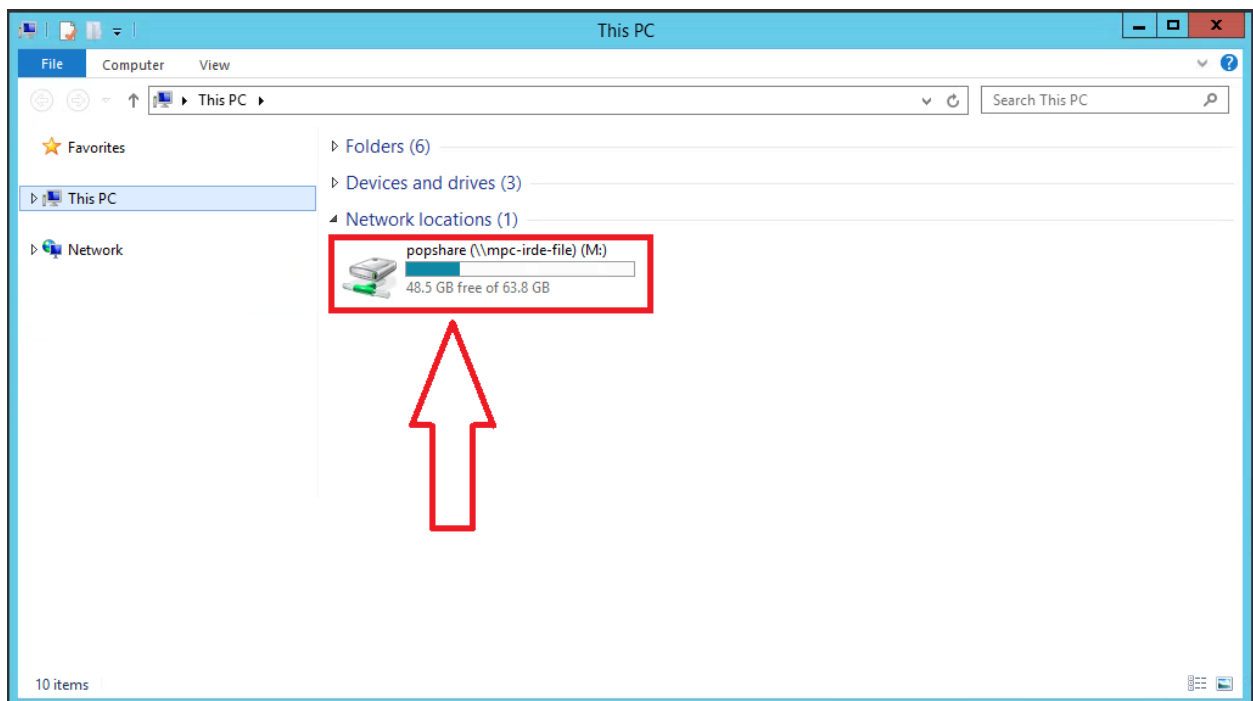
If you have a long-running computer process that you want to allow to continue while you are not using the remote desktop, you may skip the log off process and simply close the application window containing the virtual environment. However, note that you must log back in within 24 hours of closing the window or your session will be closed and your process will be interrupted if it is still running.

2.3 Navigating the file system

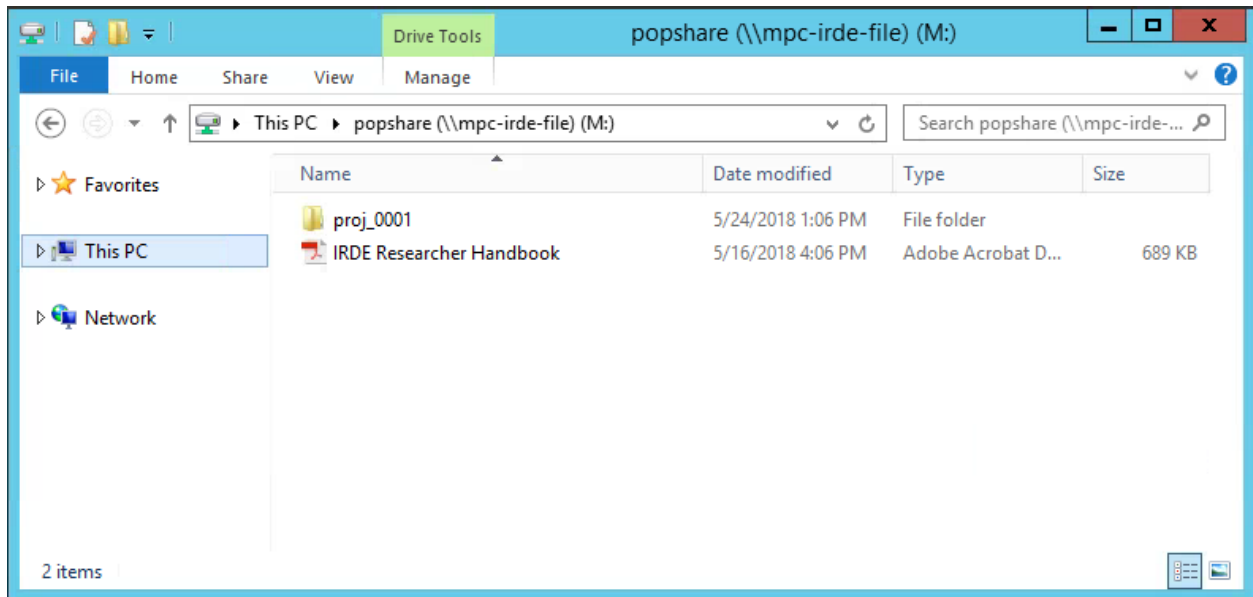
Every project has a designated folder labeled with your project ID (e.g. "proj_0001"). This is where you will find the data you requested, and this is where you should save all files you create while working in the enclave. The easiest way to access your project folder when you log in is to click the File Explorer icon in the taskbar, as seen in the screenshot below.



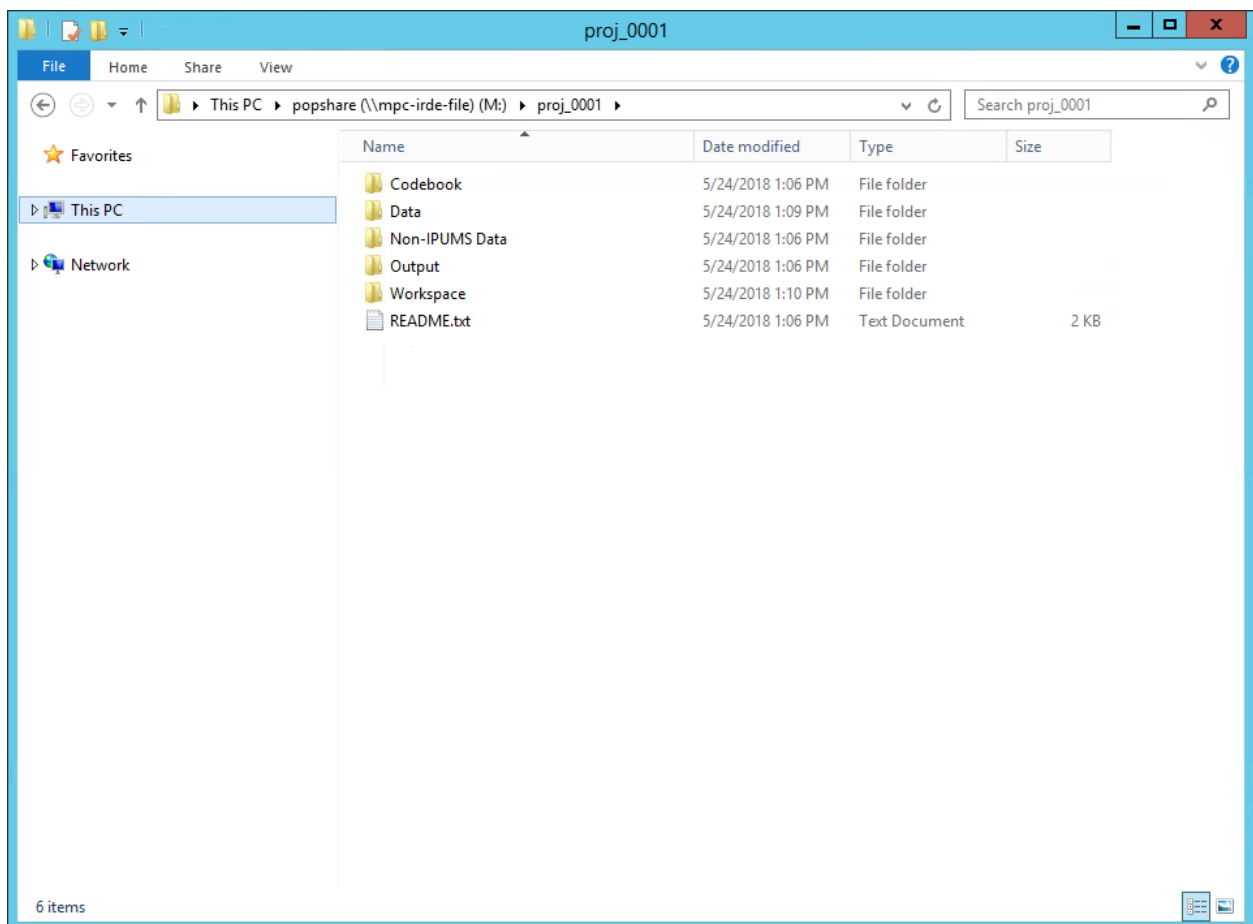
When you open File Explorer, you should see the mapped drive “popshare” under “Network locations”. Your project folder is stored on this mapped drive.



Inside popshare, you will find your project folder and a PDF copy of the IRDE Researcher Handbook.



The structure of the subfolders in your project directory is shown in the screenshot below:



The purpose and contents of these subfolders are as follows (this information is also found in the “README” text file in your project folder):

- **Codebook:** This folder contains the documentation for the IPUMS IRDE data in the Data folder. *This folder is read-only.*
- **Data:** This is where you will find the IPUMS IRDE data that you requested in your application. *This folder is read-only.*
- **Non-IPUMS Data:** This is where you will find any non-IPUMS IRDE data that you requested. If you did not request any non-IPUMS data, this folder will not appear in your project folder. *This folder is read-only.*
- **Output:** This is where you should save your final analysis output when it is ready to be reviewed by IPUMS IRDE staff. As explained in Chapter 3 of the IPUMS IRDE Researcher Handbook, you should save the files you want to be released in “To Be Released”, and save required supporting materials in “Supporting Materials”. After adding output to this folder, you should submit a request for review.
- **Workspace:** This is where you should save all your work until you are ready to submit output for review. You can copy your data to this area and create any subfolders that are helpful to organize your analysis here.

2.4 Loading the data into statistical software

Both Stata and R are installed in the IRDE virtual desktop environment. In your application for access, you can request data in fixed-width format (.dat file extension), comma-separated values format (.csv), and Stata format (.dta). Each of these formats can be loaded into either Stata or R as described in the sections below.

If you copy files from the Data subfolder to a location in the Workspace subfolder, make sure you change your file paths accordingly when following these instructions. Also, make sure that you copy all necessary files to the same location (e.g., both the .do and .dat file for reading into Stata, or both the .xml and .dat or .csv file for reading into R).

If you have not copied your data to a new location in the Workspace subfolder, the path to your data and .do files will be of the form `M:/[project_folder]/Data`, otherwise it will be of the form `M:/[project_folder]/Workspace/path/to/copied/data`.

Note that if you want to save a copy of your data after opening it in Stata or R, you will have to specify a save path in the Workspace subfolder, because the Data subfolder is read-only.

2.4.1 Stata

Before executing any of the Stata instructions, first change your Stata working directory to be your project folder by opening the File menu and choosing “Change working directory...”, clicking on This PC -> popshare -> [project folder], and clicking OK, or by running the command `cd "M:\[project_folder]"` in the Stata Command pane, filling in the name of your project folder.

Fixed-width (.dat). To read a fixed-width data file into Stata, you need to use the provided .do file. The .do file assumes that your working directory contains both the .do and .dat file, so you’ll need to change your working directory to the location of these files, which will be the Data folder inside your project folder, unless you have copied those files elsewhere.

The .do file also assumes that the name of the .dat file has not been changed, so if you change the name of the .dat file you will need to edit the name in the `using` command toward the top of the .do file.

Once you’ve changed your working directory (and updated the filename if necessary), run the .do file by opening the File menu and choosing “Do...”, navigating to the location of the .do and .dat file, and double-clicking the .do file, or by running the command `do [filename.do]` in the Stata Command pane.

Comma-separated values (.csv). *Warning: We do not provide a script to attach variable and value labels to data read into Stata from .csv files. However, variable and value labels can be found in the Codebook folder.*

In the Stata Command pane, run the command:

```
import delimited "M:\proj_0001\Data\[filename.csv]", clear
```

If you have copied the .csv file to a location outside the Data folder, make sure to modify the specified path to reflect this.

Stata (.dta). In Stata, open the File menu and choose “Open...”. Navigate to the location of the Stata data file and double-click the file to open. Alternatively, in the Stata Command pane, run the command:

```
use "M:\[project_folder]\Data\[filename.dta]", clear
```

If you have copied the .dta file to a location outside the Data folder, make sure to modify the specified path to reflect this.

2.4.2 R

Before executing any of the R instructions, first change your R working directory in RStudio by opening the Session menu and choosing Set Working Directory -> Choose Directory..., clicking on This PC -> popshare -> [project folder], and clicking OK, or by running the command `setwd("M:/[project_folder]")` in the RStudio Console pane, filling in the name of your project folder. More information on working with IPUMS value labels in R can be found by running the command `vignette("value-labels", "ipumsr")` in the RStudio Console pane after installing the `ipumsr` package.

Fixed-width (.dat) or comma-separated values (.csv). To read a fixed-width data file into R, or to attach value labels to data in a .csv file, you need to use the provided .xml file and the `ipumsr` package.

Install the `ipumsr` package by running the following command in the RStudio Console pane (note: you only need to run this command once):

```
install.packages("ipumsr")
```

To load your data, run the following commands in the RStudio Console pane:

```
library(ipumsr)
metadata <- read_ipums_ddi("Data/[filename.xml]")
data <- read_ipums_micro(metadata)
```

The `ipumsr` package assumes that your data file and .xml file are saved in the same location and have the same name (except for the file extension). If you have copied your data and .xml file to a different location in your Workspace subfolder, you will need to specify the appropriate path to the .xml file when calling the `read_ipums_ddi` function. If you have renamed either file so that their names don't match, or have stored them in separate folders, you will need to specify the path to the data file with the `data_file` argument when calling `read_ipums_micro`. Run command `?read_ipums_ddi` or `?read_ipums_micro` in the RStudio Console pane for more help with these functions.

Stata (.dta). Install the `haven` package by running the following command in the RStudio Console pane (note: you only need to run this command once):

```
install.packages("haven")
```

Then, in the RStudio Console pane, or in your R analysis script, run the following commands:

```
library(haven)
data <- read_dta("Data/[filename.dta]")
```

If you have copied the .dta file to a location outside the Data folder, make sure to modify the specified path to reflect this.

2.5 Using Stata and R packages

You may request particular Stata and R add-on packages in your *Application to Use Restricted Data*. You will be unable to download packages from the web while using the IRDE, so IRDE staff will save the package files to a shared drive accessible from within the virtual environment.

Stata packages you requested will be immediately available when you open Stata.

To use an R package you requested, you must first install the package with command:

```
install.packages("[package_name]")
```

R has been configured to install packages from a local package repository. Once you've installed the package, you can access its functions using the `library([package_name])` function or the `[package_name]::` prefix.

Chapter 3

Reviewing output for disclosure risk

Before analysis output can be released to you for use outside the virtual environment, IRDE staff must review it and determine that it abides by the rules described in this chapter. Moreover, you are responsible for knowing the disclosure rules described here, and for reviewing your own output to ensure that it follows these rules.

3.1 Disclosure protection rules

Broadly speaking, all output must meet two criteria to be cleared for release from the IRDE virtual environment:

1. The output must not reveal any confidential information about any particular household or individual.
2. The output must not reveal that any particular household or individual is present in the IRDE dataset.

The second criterion would be a concern if you were to combine IRDE data with outside datasets that contain identifying information. IRDE staff will ensure this criterion is met by reviewing any outside data which you request be added to your enclave workspace. Outside datasets with identifying information will not be permitted.

To determine whether output meets the first criterion, and does not reveal any confidential information, the IRDE uses a *threshold rule* and a *concentration rule*, depending on the type of output.

3.1.1 Threshold rule

A *threshold rule* states that any output cell containing, or derived from, a number of observations below a certain threshold is considered a disclosure. This rule applies to tabular output as well as output from regression models that include categorical variables. The IRDE does not publicize our small cell thresholds, but IRDE staff will inform you of applicable thresholds as necessary during the output review process. More detail on how we handle small cells is provided below in the section on modifying disclosive output.

3.1.2 Concentration rule

A *concentration rule* is most often applicable to economic magnitude data, such as aggregated sales figures for firms by industry or sector. Because IPUMS IRDE data focus on households, concentration rules will rarely be applied, but there may be some situations where they are relevant.

In general terms, a concentration rule states that the value of a cell must not be dominated by a few cases. Such a rule is operationalized using either a $p\%$ rule or an (n,k) rule.

A $p\%$ rule specifies that the case with the second largest value in a cell should not be able to determine the value of the case with the largest value within $p\%$. For instance, if the aggregate sales value is \$100 million, and the second largest firm has sales of \$40 million, that firm can deduce that the largest firm has sales of approximately \$60 million. If the true largest value is \$50 million, the second largest firm's estimate is within $\$10 / \$50 = 20\%$ of the true value. Thus, this cell would be considered too concentrated for values of $p \geq 20$ or greater. The value of p used by the IPUMS IRDE is kept confidential, but will be provided to you if applicable once you have been granted access.

An (n,k) rule specifies that the n largest cases in a cell must not contribute more than $k\%$ of the cell total. The values of n and k used by the IPUMS IRDE are kept confidential, but will be provided to you if applicable once you have been granted access.

3.2 Preparing output for review

Typically, the analyses you perform in the IRDE will form the basis of a scholarly product, such as a paper or presentation. Ideally, you should submit only one request for output review per paper or presentation (or revision thereof), when you have generated all output needed for that product.

At this time, you should put *the exact files you want released* in the “Output/To Be Released” folder. You should put additional materials that will help IRDE staff review the output you want released in the “Output/Supporting Materials” folder.

Examples of supporting materials include the analysis scripts used to generate the output to be released, and additional tabulations that show the cell counts underlying the output to be released.

Finally, you should include a plain-text file named “README.txt” in “Output/Supporting Materials” summarizing the contents of the files in “To Be Released” and “Supporting Materials”, including which supporting materials pertain to which pieces of output.

3.2.1 Tabular output

You should strive to request as little tabular output as necessary, as information from multiple tables with overlapping variables or samples can be combined, creating a high risk of disclosure. Never request review of tabular output until you are sure you have the table specification exactly as you want it, because once a table has been released, versions of the same table with minor modifications may pose too high a risk of disclosure to be released.

In the case of sample data, cell counts in tables to be released should generally be weighted with sample weights, though unweighted totals may be allowed if necessary for estimation of statistical significance. You must include the corresponding unweighted tabulations in the “Supporting Materials” folder, so that IRDE staff can easily see the number of observations underlying the weighted table cells.

One way to expedite review of tabular output is to proactively combine cells with few observations, especially when distinctions between such cells are not important to your research question.

3.2.2 Summary statistics

Disclosure rules for summary statistics follow the same logic as those for tabular output. Minimum and maximum values at the level of individuals cannot be released because they are based on a single observation. However, minimum and maximum values – and other summary statistics such as means, medians, standard deviations – can be released at levels of aggregation above the individual as long as the underlying number of

observations exceeds the cell size threshold. For instance, the minimum and maximum rates of employment by province can be released if the number of observations from the relevant provinces exceed the cell size threshold.

As with tabular output, summary statistics to be released from sample data should generally be weighted, and in all cases, you must include the unweighted cell counts underlying the summary statistics in the “Supporting Materials” folder.

3.2.3 Regression output

Coefficients from a regression involving only continuous variables pose little risk of disclosure, as long as the sample size exceeds cell size thresholds. For regression output with only continuous variables, it is sufficient to document the number of observations on which the regression is based, being sure to take into account observations that were dropped from the analysis due to missing values.

If your regression analysis contains any categorical variables, either as predictor or response, you must document (in the “Supporting Materials” folder) the unweighted number of cases included in the regression in each category of that variable. If you have both a categorical response and one or more categorical predictors, you must document the unweighted number of cases included in the regression that fall into each cell of a cross-tabulation between each categorical predictor and the response variable. Similarly, if your regression includes interactions between categorical predictors, you must document the cell counts of those cross-classifications.

The only exception to these rules is for categorical predictors included in your model for which you do not need to release coefficients. If you are not requesting release of coefficients for a predictor, you do not need to document the underlying cell sizes. Suppressing coefficients allows you to control for factors such as low-level geography without creating undue risk of disclosure.

3.2.4 Graphical output

Graphical output in which points represent individual observations cannot be released. Any point estimate displayed in a graph must be based on a number of observations that exceeds the cell size threshold, and the unweighted number of observations underlying each point must be documented in “Supporting Materials”.

Point smoothing techniques may be used in cases where the number of observations underlying point estimates is small. For instance, histograms may not be released, but kernel density plots may be allowed, though the distribution’s tails may need to be suppressed.

3.3 Modifying disclosive output

It is best to proactively modify output that may be disclosive before requesting review, such as by combining categories from tables or regression predictors with small numbers of observations. If you submit output for review and IRDE staff determines that some of that output must be modified, they will most likely suggest that you combine small cells to reduce the risk of disclosure. In the case of tabular output, combining small cells is almost always preferable to cell suppression, because cell suppression requires complementary suppression of additional cells, and this process is complex and will delay the release of output.

3.4 Requesting review of output for release

Before requesting a review of output, make sure your answer to each question below is “Yes”:

- Have I generated all of the output that I need from the IRDE for the scholarly product (e.g., paper or presentation) on which I am currently working?
- Have I saved all of the output I want released in “Output/To Be Released”?
- Have I saved all required supporting materials in “Output/Supporting Materials”?
- Have I created a plain-text file named “README.txt” which describes the output to be released and the corresponding supporting materials, and saved this file in “Output/Supporting Materials”?

If the answer is “Yes” to each of these questions, send an email from the address you used to register with the IRDE to ipums-irde@umn.edu requesting review, and be sure to include your IRDE username.

3.5 Output release procedure

Depending on the volume and complexity of your output, it may take IRDE staff up to five business days to review your request for release. If IRDE staff determines that your output does not pose undue risk of disclosure, they will email your output to the address associated with your IRDE account. If your output cannot be released without modification, IRDE staff will send you an email describing why the output cannot be released in its current form, and suggesting modifications that would enable the output to be released.

Chapter 4

Concluding a project

4.1 Archiving project files

The IPUMS IRDE will securely store all of the files from your project directory for five years after the conclusion of the project. At that time, we will contact you to inform you that your files are scheduled for deletion, and offer you the chance to request a storage extension if you intend to return to your analysis in the near future.

4.2 Citing the IPUMS IRDE

Please cite the IPUMS IRDE in all publicly presented and published works. The preferred citation is below. Cleveland, Lara, Matt Sobek, and Steven Ruggles. 2018. *IPUMS International Research Data Enclave*. Institute for Social Research and Data Innovation, University of Minnesota.