Agreement and instructions on how to use the IPUMS/ancestry.com full count census records on the demography/Sociology/CEDA/PopCenter network

April 24, 2017

Contents

1 NEW Features in 2017

Several new features have been added and a several procedures have been simplified in 2017. You should read this section even if you have been using the full count IPUMS data since high school.

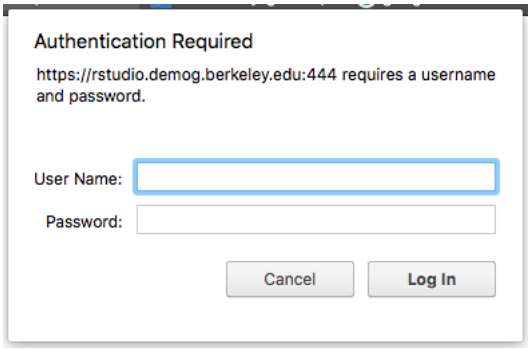
- The data and all '-ipums' user accounts have been moved from `keyfitz` to `secureIpums`. So where you used to type things like `ssh carlm-ipums@keyfitz` you will now type things like:

```
carlm@quigley:~$ ssh carlm-ipums@secureIpums
```

- The new host, `secureIpums`, is a virtual machine and only authorized users have access to it. Consequently, the encfs encryption scheme for the ipums-repo data is no longer necessary. When you logon, you should have immediate access to the ~/IPUMS directory, in which all the full count data are stored. If this is not the case, then the following Linux command should fix it:

```
carlm-ipums@secureIpums:~$ ln -s /ipums-repo2 ~/IPUMS
```

- Harmonized IPUMS data for 1920 and 1930 are now available in addition to the 1940 data. For other census year, we still have only the Ancestry.com .txt files.
- In order to minimize waste, fraud and abuse the 1920, 1930 and 1940 data, are provided in both STATA .dta and .csv formats.
  - In census year subdirectory, e.g. ~/`IPUMS`/1930 there (at least) three subdirectories:
    1. **Ancestry** holds the .txt files from Ancestry.com
    2. **CSV** holds the .csv files which are designed to be used with R's blazingly fast `data.table::fread()` command. See the README files in the CSV directories for details and examples on how to make complex extractions in the blink of an eye.
    3. **DTA** holds the STATA .dta files. Look for README files in the DTA directory for details on how you might use these files.
    4. **MPC** holds the original files from Minnesota Population Center. You should not need to access these files unless you crave suffering.
- Rstudio access is now possible without first logging into noMachine. To get Rstudio in your browser the following is required:
  - you must be added to a special group (ask carlm)
  - you must browse to `http://fc.demog.berkeley.edu`
  - you must encounter a dialog box like this one:



    And reply with your standard Demography Lab userid and password (what you would use to log into noMachine).
  - After clearing the above described dialog box, you should see the familiar Rstudio login screen. Here you use your 'user-ipums' userid and password.

2 Overview

The Demography Lab network houses secure **nonpublic** census records for census years 1790 - 1940. Access to these data are granted by agreement between UC Berkeley, and the Minnesota Population Center (IPUMS USA). Contact Josh Goldstein if you have not already secured permission to use these data.

Because these data are both proprietary and "big" using them requires that you agree to abide by our contract with the Minnesota Population Center and tolerate a few inconveniences thereby entailed:

- The data can only be accessed from specially enabled computer accounts on the machine called `secureIpums` which cannot be accessed directly from machines that are **not** on the Demography Lab network. You **can** still work with these data remotely, however, See Section 4.1 for details.
- We all need to be mindful of disk use. If we all keep lots of private versions of these huge datasets, we will fill up our disks and ruin it for all of us.
- The data and the files that you create in your special account are only **automatically** backed up in a limited way that is intended to protect your program files (which are valuable) while excluding data files (which are sensitive and easily reproduced). See Section 4.5 for details. Please do not act in a way that will subvert this goal.

3 Initialization, setup, configuration

By necessity, the first step you must take is to fill out some forms, and download, install and configure some software. This stuff only needs to be done once, so close your eyes and think of England.

1. **Get permission from Minnesota Population Center to use the data** You've probably already cleared this hurdle, but if not, email Carl at cmasonberkeley.edu for instructions on which forms to fill out.
2. **Establish a standard Demography Lab account** by filling out and returning the one page *Statement of Compliance*

   `http://lab.demog.berkeley.edu/Docs/statementofcompliance.pdf`

   Email or otherwise send it to cmason@berkeley.edu. Obviously, this is only necessary if you do not already have a standard Demography Lab account.
3. **Establish a second `secureIpums`-only account** by sending email to cmason@berkeley.edu – who will ask that you sign the last page of this document indicating your acquiescence to the security protocol
4. **Install and configure the appropriate freeNX client** visit
   `http://lab.demog.berkeley.edu/LabWiki/index.php/Help:Contents`
   and follow the instructions regarding the freeNX client. As of this writing the *NoMachine* client is the best but this could change.

That's it. At this point you have all the tools you need to do science.

4 Best practices

4.1 (Remote) access to the computing environment and `secureIpums`

Access to `secureIpums` can be achieved either from the workstations in the 2232/2224 Piedmont, or remotely via a freeNX connection. (It is also now possible to access the IPUMS data in Rstudio, see Section 1 for instructions.) Regardless of the method used, gaining access to the full count data requires **two distinct login accounts and passwords**. (A third password to decrypt the data **is no longer required**.) After completing the login procedure described below, you will be able to operate in the following manner:

You will have a "(virtual) desktop" running either on `quigley` (if you logged in via freeNX) or `immigrant` (if you are on a demography workstation). From within the desktop, you will open a terminal window that will connect to a second machine called `secureIpums`. From within that terminal window, you will have read access to the full count census data.

As noted above, accessing system on which the data reside requires two login accounts. The first account is a standard Demography Lab account that allows you to login to workstations in 2232/2224 Piedmont or remotely using freeNX. The second account is special in that that it will only work on `secureIpums` and is enabled to access the secure data, but otherwise, it behaves just like the standard account. It will have a name that ends in '-ipums'.

Once the accounts and passwords have been acquired and the freeNX client installed and configured (See Section 3), the multi step login procedure is below:

1. **Launch your freeNX (NoMachine) client and connect to `quigley`** This should produce a live Linux desktop in a within a window on your local computer. Although I describe it as "one step" it actually requires quite a few mindless clicks on the "next" button. One gets used to it.
2. **Get a terminal window** A terminal window application can be launched from **Applications→Accessories→ROXterm**

   Inside the terminal window, ssh to `secureIpums`

   ```
   carlm@quigley:~$  ssh carlm-ipums@secureIPUMS
   ```

   of course your `secureIpums` userid isn't "carlm-ipums" is it?
3. After giving your userid-ipums password your terminal window should change slightly to reveal that you are now on the machine `secureIpums`.

The first time you ssh to `secureIpums` the machine should insist that you change your password. If it fails to do so, please use the `passwd` command in the ssh window to change your password:

```
carlm-ipums@secureIpums:~$ passwd
```

That's `passwd` not passwORd, and don't be put off by Linux unusual failure to echo your passwords as '***'. One imagines that enemy agents lurk everywhere... which of course means that you should change your password often.

The ssh window on `secureIpums`, with your userid-ipums account running in it, is where you can do science.

xstata and others You do not need to do all your work from the Linux prompt. Stata, R, Rstudio, Python, and possibly other tools if you need them, are available. The only catch is that you need to launch them from the Linux prompt because you do not have access to an Application menu on `secureIpums`. For gui interface STATA, simply cd to an intersting directory and then type:

```
carlm-ipums@secureIpums:~$ xstata-mp
```

To run R, you probably want to first launch emacs, but you can also use Rstudio in your browser (See Section 1).

Directory structure There is a directory in ~/IPUMS for each census year for which we have data. Each directory is organized somewhat differently, but generally each directory contains the data in several formats: .csv for blazing fast reads in R (using data.table::fread) and .dta files for STATA. Look for helpful README.txt files in the relevant directories.

4.2 Writing intermediate files

There are three important issues to consider regarding the creation of intermediate files.

All sensitive data must stay on secureIpums First and most important, In order to keep our promises it is essential that the raw data must never leave secureIpums. It would be hard to make this happen accidentally, but be careful. In order to protect ourselves from accidents, /90days and /data/commons are accessible as read only. They are useful therefore, for bringing data into secureIpums but not the other way. See section 4.3 for instructions on extracting results.

Disk space limitations Second, since the data files are huge and disk space is a shared and limited resource, it is important that we avoid making unnecessary copies of stuff. To this end, the 1920, 1930 and 1940 data are stored in several different ways. Check the README files in the CSV, and DTA directories before just creating your own scale model of the universe.

Obviously, it would be nice also, if you could delete old copies of stuff that no longer serve a purpose.

There is no need to worry about zipping or compressing files, however, as this is done automatically at the filesystem level. That is, everything is compressed and uncompressed as it is written and read from the disk. Because computers are faster than disk drives, this turns out to be good idea now days. Zipping and unzipping files on secureIpums is therefore, entirely counterproductive.

Secret temporary files Third, programs such as STATA store temporary files in places that you might not be aware of. In the case of STATA, your account *should* be configured to store them in either Private (if your account was created before 2017) or STATATEMP otherwise. R tends to want to make a copy of your environment in .RData whenever you exit the program. In both of these cases you can wind up using lots more disk space than you realize. As long as R or STATA is not presently running, it is perfectly safe and useful to remove such detritus.

```
carlm-ipums@secureIpums:~$ /bin/rm  /.RData
```

or

```
carlm-ipums@secureIpums:~$ /bin/rm ${STATATMP}/*
```

A bigger *potential* problem is that many programs will by default write temporary files to system partitions that can fill up and crash the machine. If you use software other than R or STATA Please take the time to figure out where it puts its junk and make sure, if you can, that it writes it into your home directory–where you can clean it up from time to time.

4.3 Repatriating results

When you make a scientific breakthrough, it is nice to be able to share it with the world. To do so, you will want to copy results (and of course **only** the results) to your Demography Lab account and from there to snapchat or wherever.

The easiest way to do this is with `scp`. To move a file called `cure-for-cancer.txt` from your carlm-ipums@`secureIpums` account to your Demography Lab account (carlm) you could type the following:

```
carlm-ipums@secureIpums:~$ scp cure-for-cancer.txt carlm@nmx.demog.berkeley.edu:
```

If you need to move a bunch of files, it is efficient to either create an archive file using `tar` or `zip` or to use `sftp` to move several files at a time. To launch sftp type:

```
carlm-ipums@secureIpums:~$ sftp carlm@nmx.demog.berkeley.edu
```

Then type '?' at the resulting `sftp>` prompt to see what can be done.

Note that in both examples you are logged into your user-ipums@secureIpums account and are transferfing files from there to your Demography Lab account.

4.4 Disconnecting and logging off

It is important to logoff of the system when you are finished working not only to protect the data – from those who would sit in your still warm seat and do grave mischeif– but also because licenses and RAM are all shared. Leaving STATA running with 300GB of RAM while you are sleeping, is likely to slow human progress. Yes, it can be boring to have to read in those big files *again*, but some of your colleagues are nocturnal and can really benefit from the RAM that you are not using.

4.5 Backups

Because secure data ought not to be stored in the cloud (where we keep our normal backups) we will not be backing up your `secureIpums`-only account comprehensively. Instead, we will make backup copies of all files in your `secureIpums`-only account that are *under 10MB in size*. The idea is that very few code files are ever that large, whereas very few data files are ever that small.

This plan will cover you in case of normal hardware disasters, but bear in mind that we will probably not be able to restore files that were accidentally deleted a week ago, nor will we be able to recover your work after a fire guts our machine room. It would not be a bad idea to use a version control system such mercurial if you are concerned about being able to reproduce past versions of your code. And consider developing a plan to routinely save your code files to your Demogrpahy Lab account–which is backed up the cloud. Contact cmason@berkeley.edu for help with this sort of thing.

ALSO note that – if your account was created before 2017 – we cannot backup anything in your `Private` folder. Or more to the point, we cannot backup anything from that folder in such a way that it could ever be decrypted later. Consequently, it is best practice to keep you program files in a separate non-encrypted directory and perhaps tar and scp it to your Demography Lab account from time to time. (See Section 4.3 for instructions on how to get stuff from `secureIpums` to `quigley`.

5 Agreement

I have read this entire boring *Agreement and instructions on how to use the IPUMS/ancestry.com full count census records on the Demography Lab network* and enthusiastically agree to uphold the terms of the understanding between UC Berkeley and the Minnesota Population Center by following all of the guidelines described herein.

_____

(signature)