

dEntry: Digitization of Historical Records through Crowdsourcing

Roy Mill

November 23, 2012

Abstract

This document describes a web system I developed for the facilitation of the digitization of historical censuses by data entry clerks hired through an online labor market. The system contains a convenient interface for clerks to enter handwritten records and a queue that allocates records to clerks and dynamically determines double and triple entry according to agreement or disagreement between clerks. The system can also track employee quality through the comparison of clerks' entered data. Having direct link to the database, records that have been entered can be pulled to statistical analysis software (e.g., Stata) at any moment.

I hire about 30 Data Entry Associates through a leading online labor market.¹ I then refer them to dEntry. On dEntry, data entry clerks get images of scanned census sheets at the top of the screen and a form that matches the census form at the bottom of the page. Clerks need to enter the information they read in the sheet into the form. The interface that clerks use can be seen in the training video that I refer them to before they start working: http://www.youtube.com/watch?v=58abfly_ieI.

While clerks are entering data, and even before they start entering, several things happen in the background:

1. We start with data from FamilySearch.org that has some basic details from the form already filled out (names, ages, races, etc), as well as metadata on what roll, image, and row the records in question should appear. A command in Stata that I wrote uploads these records to the dEntry system along with a priority number that determines the order of the queue of records to enter.
2. When clerks log in they get the next household from the aforementioned queue. After saving the data for the current household the system decides whether to send the same household to another clerk to enter while the clerk who just saved the household moves on to the next one. If the record was randomized to be re-entered by someone else, then we compare the entered information after the second clerk has finished entering. If the two clerks do not enter the same data for the same records, the household will be automatically given to a third clerk.

¹Some examples for such websites are freelancer.com, odesk.com, and elance.com.

3. Double (and triple) entry helps to detect errors but it also doubles and triples costs. We dynamically set the share of records that are re-entered. We then set a probability $p_r \in [p, 1]$ for which records will be sent for re-entry, with an initial probability $p_0 = 1$.
 - (a) The first batch of records for a starting clerk will be given with probability p_0 to another clerk .
 - (b) Whenever the clerk agrees with another clerk on a record, this probability is reduced by $d < 1$, and the next batch of records will have $p_{r+1} = \max \{dp_r, p\}$.
 - (c) When there is disagreement between the first two clerks and then a third clerk is asked to enter the data, if the third clerk agrees with one of the first two, the clerks who agree will have their probability decreased by d and the one who disagreed with them will have his probability increased by $u > 1$ to be $p_{r+1} = \min \{up_r, 1\}$.
4. After enough batches, p_r becomes a measure of how inaccurate a clerk is. It relies on the assumption that errors do not tend to be the same (and that coordinating on mistakes is too costly). We use this to screen out freelancers who are not accurate and to keep track of the current ones.
5. When we import the data into Stata there is still some cleaning to be done (not much usually), and most importantly part of the import command that I wrote for Stata will also determine which version to use if more than one exists for a record in question. We take the value that the majority of clerks have entered. If there is no majority (each clerk entered something different) then the version of the clerk with the best record (in terms of agreement with other clerks) will be taken.

The data that clerks enter and save can be imported into Stata at any moment, so even when only some of the records have been entered, I can already start analyzing them.