Record Indexer Overview

Introduction

To assist family history researchers in finding information about their ancestors, the LDS Church is assembling a vast collection of historical records from across the world. Some of these records have been obtained from governmental organizations (e.g., census, immigration, military, tax, property, birth, marriage, death, and cemetery records). Other records have been obtained from religious organizations (e.g., christening, baptism, birth, marriage, and death records).

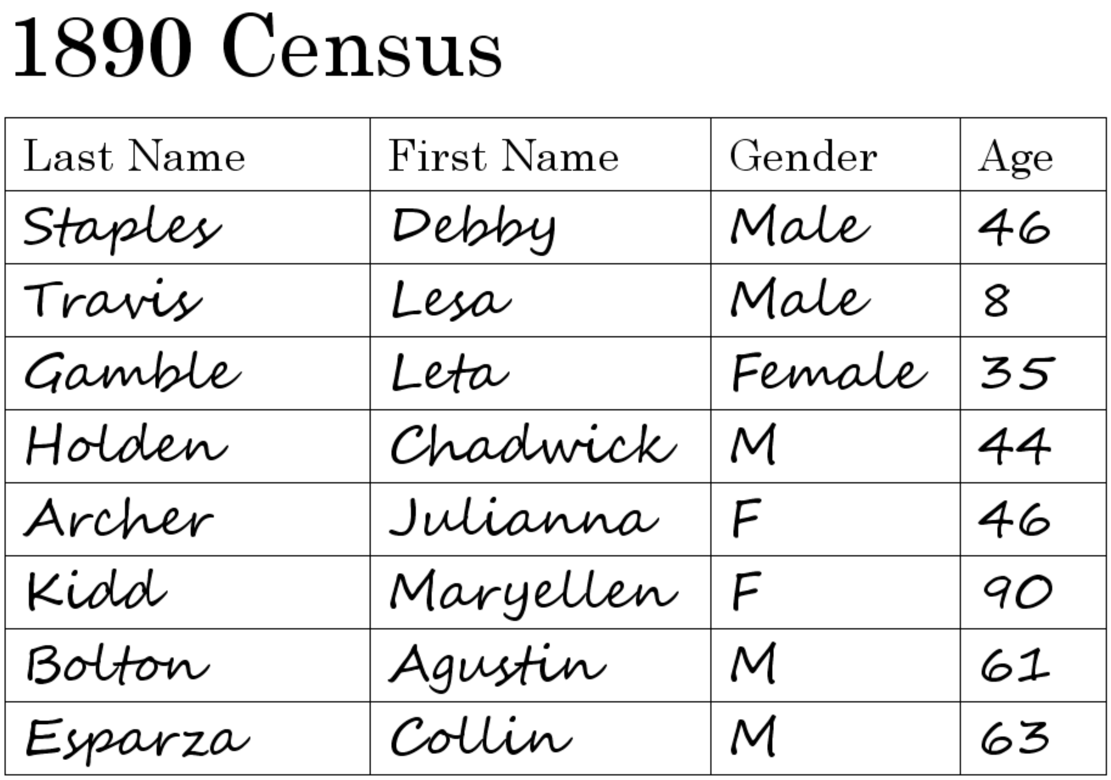
Many of the records collected by the Church are in the form of scanned images. Although the images contain names, dates, and places of interest to family history researchers, it is difficult to locate relevant images and the data they contain because the text on the images is not easily searchable. This is because scanned images are stored as two-dimensional arrays of pixels, where each pixel has a certain color. While people are good at reading images and extracting text from them, computers are still not very good at doing so (although they are getting better). Because images are stored as two-dimensional arrays of colors rather than as text, it is hard to search the images for key words because computers don't understand the text in the images.

To make it possible for family history researchers to perform key word searches on its collection of historical records, the Church has built an online system that allows volunteers to help manually "index" the images. To "index" an image, a volunteer downloads the image, reads all of the names, dates, places, and other information on the image, and types the information into the system in a textual format. The system then assembles a large database that remembers all words found on each record image. This allows family history researchers to locate records images containing relevant information by doing key word searches.

The Record Indexer project for CS 240 is a simplified version of the Church's online indexing system. The project consists of a "Server" and a "Client". The Server is the back-end part of the system that keeps track of all of the record images that need to be indexed, downloads images to volunteers for indexing, stores the words found on each image in a database, and allows the images to be searched by key word. The Client is the part of the system that volunteers use to view the images, to type in the text they see on the images, and to send the captured text to the server for storage.

Terminology

This section introduces some terminology that is used throughout the Record Indexer project specifications. As an example, consider the following image, which is part of the 1890 U.S. Census. (Obviously, we are using fake images for this project.)



1890 Census

Last Name	First Name	Gender	Age
Staples	Debby	Male	46
Travis	Lesa	Male	8
Gamble	Leta	Female	35
Holden	Chadwick	M	44
Archer	Julianna	F	46
Kidd	Maryellen	F	90
Bolton	Agustin	M	61
Esparza	Collin	M	63

A **Project** is a collection of images. Each image contains a table of data values that need to be indexed. Example Projects are “1890 U.S. Census”, “1900 U.S. Census”, “Ellis Island Immigration Records”, and “1967 U.S. Military Draft Records”. Each of these Projects consists of many images that need to be indexed.

Each column in the table represents a particular **Field**. The Fields in the example image are: Last Name, First Name, Gender, and Age. The first row in the table contains the names of the Fields.

Each row in the table (except the first) represents a **Record**. A Record is a collection of related field values. Typically, a Record contains information about one person. For example, the last record in the example image contains information about a person named Collin Esparza who is a 63 year old male.

It is assumed that all images in a project have the same structure (i.e., the same fields in the same order and the same number of records).

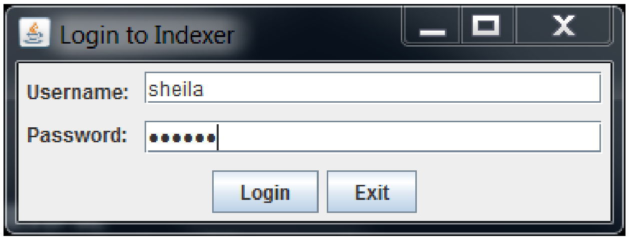
Each image is called a **Batch**, because each image contains a "batch" of one or more records. Therefore, a Project is a collection of Batches. Each Batch contains a list of Records. Each Record contains a list of Field values.

Demo

This section demos the Record Indexer system from the end-user perspective. The intent of this demo is to provide a high-level overview of how the Record Indexer system operates without going into excruciating detail. Further detail may be found in the Record Indexer Server and Record Indexer Client project specifications.

User Login

When a user runs the Record Indexer client, they are asked to login (Figure 1).

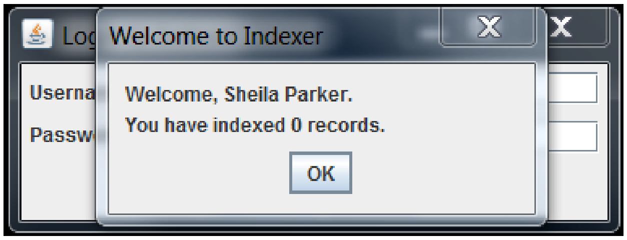


Username: sheila

Password: ••••••

Figure 1 – Login Dialog

After providing their username and password, the user is shown a welcome dialog that includes their name and the total number of records they have indexed since creating their account (Figure 2).



Welcome, Sheila Parker.

You have indexed 0 records.

Figure 2 – Welcome Dialog

After closing the welcome dialog, the main indexing window appears (Figure 3). If the user was not working on indexing a batch the last time they logged out, the indexing window appears empty, as shown in Figure 3. However, if the user was in the process of indexing a batch when they left, the state of the Client is restored to the state it was in when they logged out.

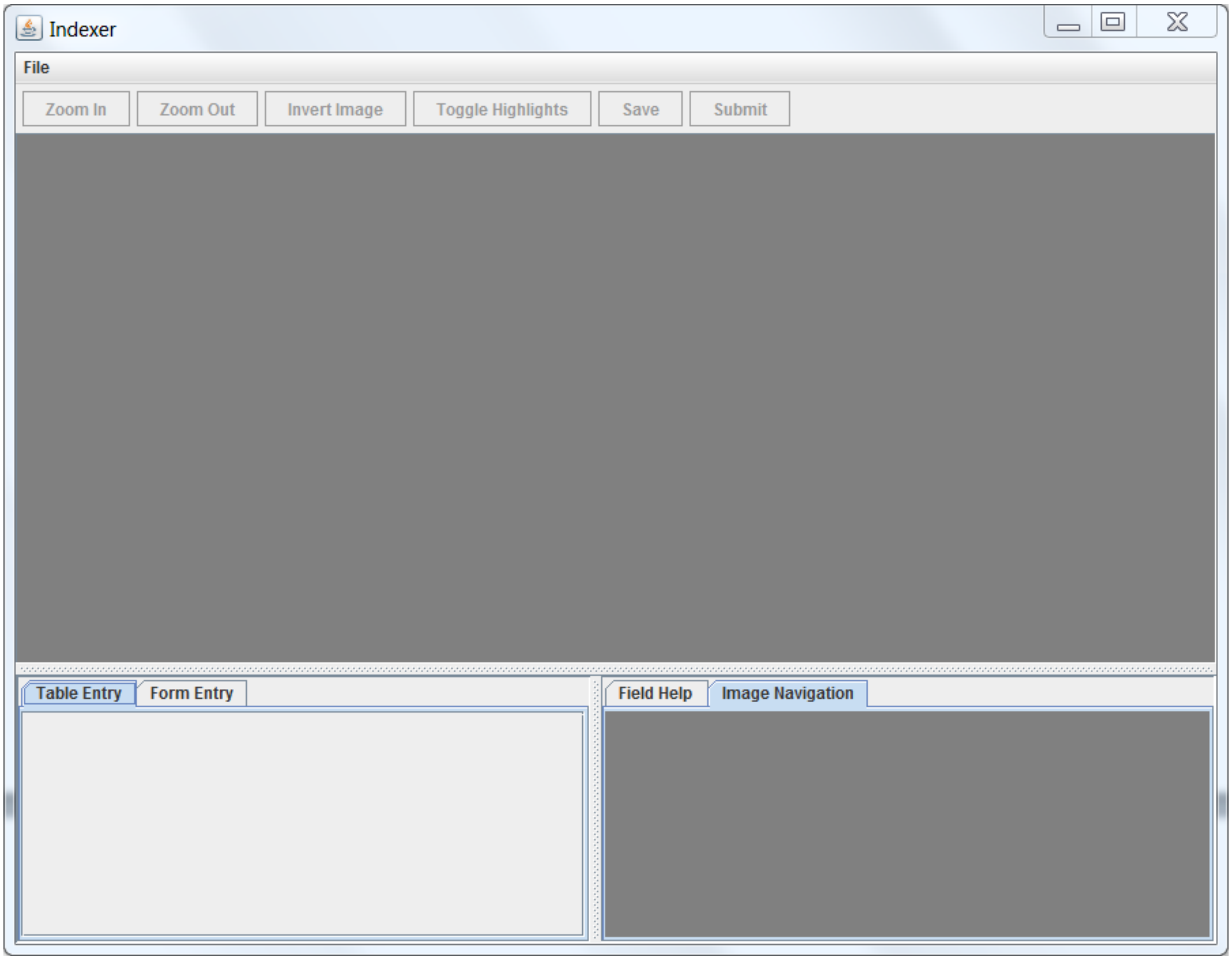


Figure 3 - Indexing Window

Download Batch

After logging in, the user can request a batch to index by selecting the "Download Batch" option from the File menu (Figure 4).

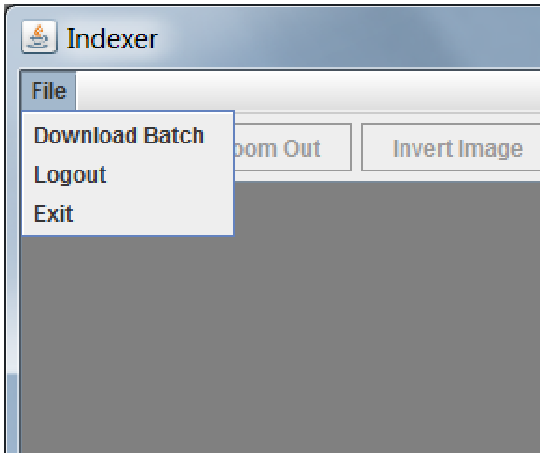


Figure 4 - File Menu

After selecting "Download Batch", the download batch dialog appears (Figure 5).

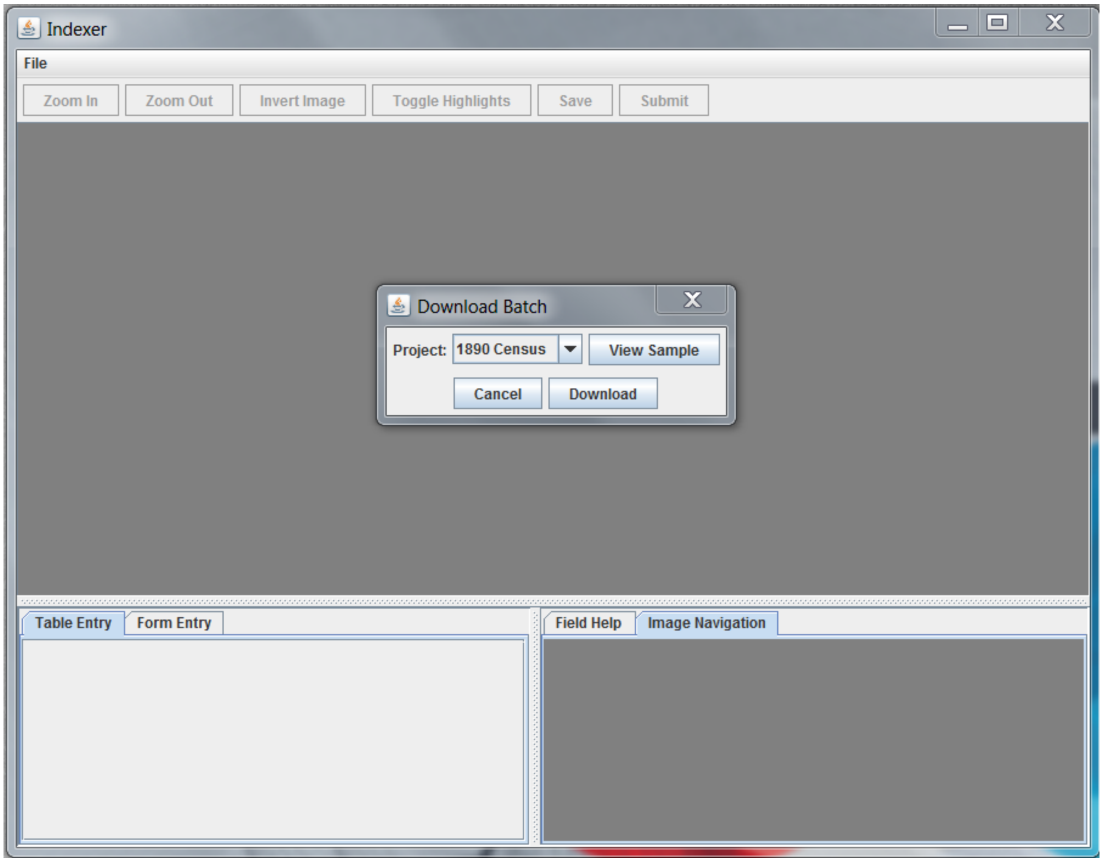


Figure 5 - Download Batch Dialog

The download batch dialog displays a drop-down list of all projects that the user can choose from. In Figure 6, the choices are “1890 Census”, “1900 Census”, and “Draft Records”. Each of these projects has batches that need to be indexed, and the user can select whichever project they prefer.

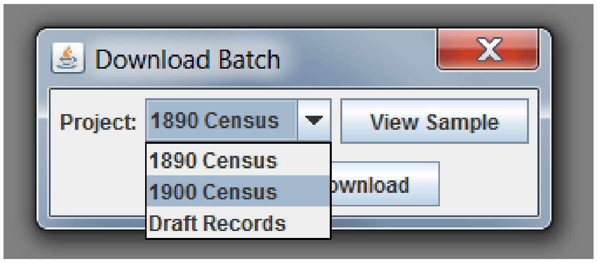
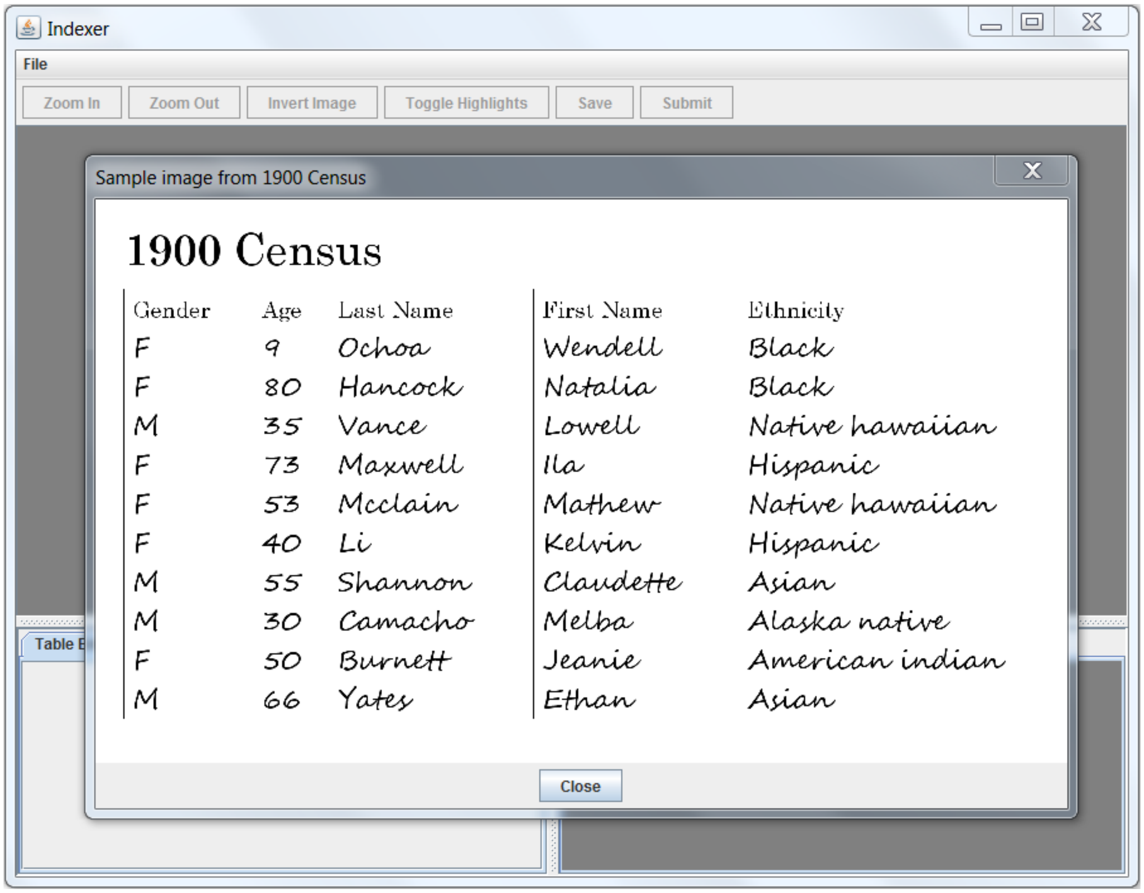


Figure 6 - Project List

If the user is not sure which project they want to work on, they can click the “View Sample” button to see a sample image for the currently-selected project. For example, if the user selects “1900 Census” in the project list and then clicks “View Sample”, they are shown a sample image from the “1900 Census” project (Figure 7). Hopefully, seeing a sample image will help them decide whether or not they want to work on this project (people typically like to work on projects for which they are able to decipher the handwriting.)

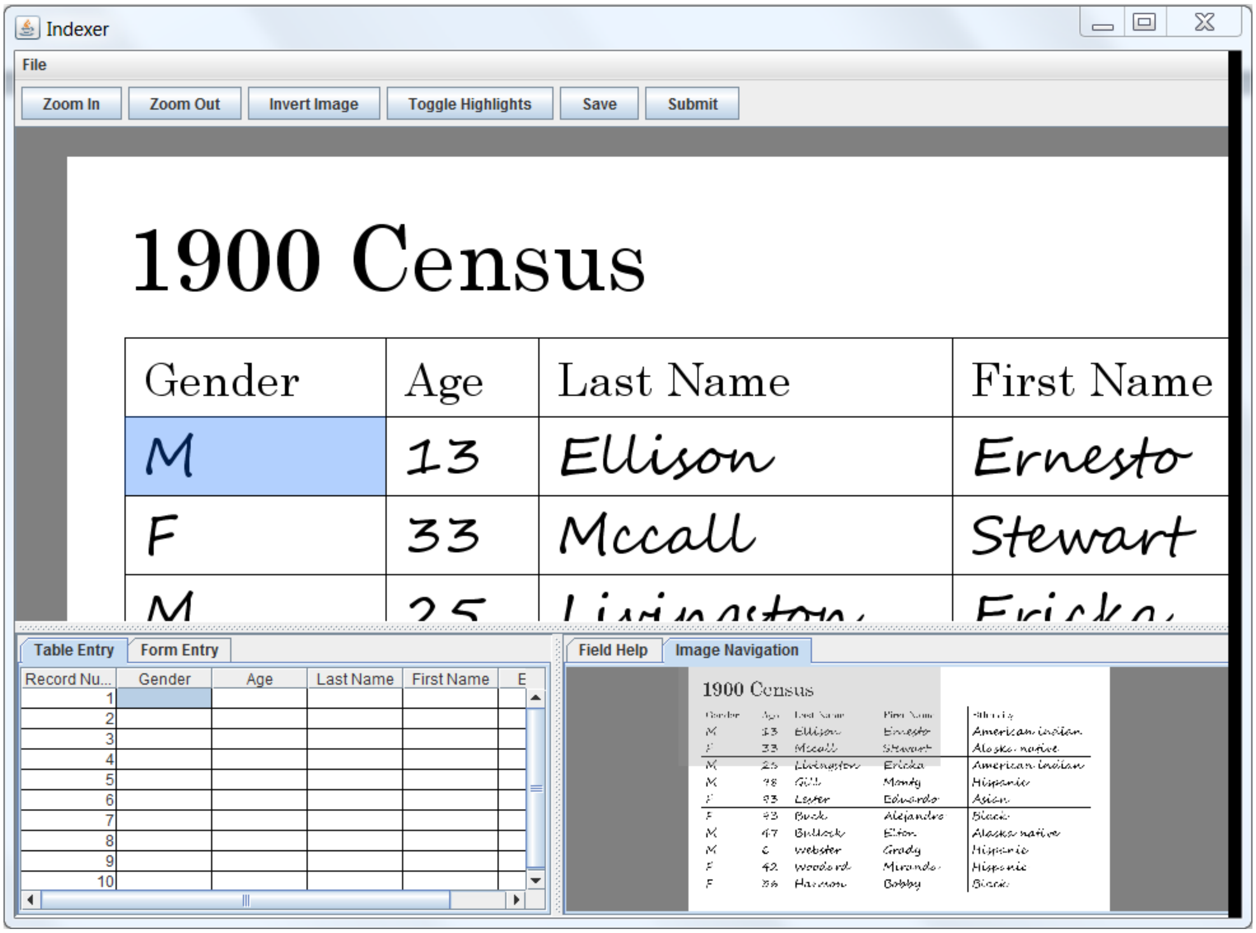


1900 Census

Gender	Age	Last Name	First Name	Ethnicity
F	9	Ochoa	Wendell	Black
F	80	Hancock	Natalia	Black
M	35	Vance	Lowell	Native hawaiian
F	73	Maxwell	Ila	Hispanic
F	53	Mcclain	Mathew	Native hawaiian
F	40	Li	Kelvin	Hispanic
M	55	Shannon	Claudette	Asian
M	30	Camacho	Melba	Alaska native
F	50	Burnett	Jeanie	American indian
M	66	Yates	Ethan	Asian

Figure 7 - Sample Project Image

Once the user decides which project they want to work on, they click the “Download” button in the download batch dialog. At this point the server will select a batch from the selected project and assign it to the user. As long as the user is working on the batch, it will not be assigned to any other users. The batch assigned to the user is downloaded to the client and displayed in the indexing window (Figure 8).



1900 Census

Gender	Age	Last Name	First Name	Ethnicity
M	13	Ellison	Ernesto	American indian
F	33	Mccall	Stewart	Alaska native
M	25	Livingston	Ericka	American indian
M	78	Gill	Monty	Hispanic
F	93	Lester	Edwardo	Asian
F	93	Buck	Alejandro	Black
M	47	Bullock	Elton	Alaska native
M	6	webster	Grady	Hispanic
F	42	Woodard	Miranda	Hispanic
F	56	Harmon	Bobby	Black

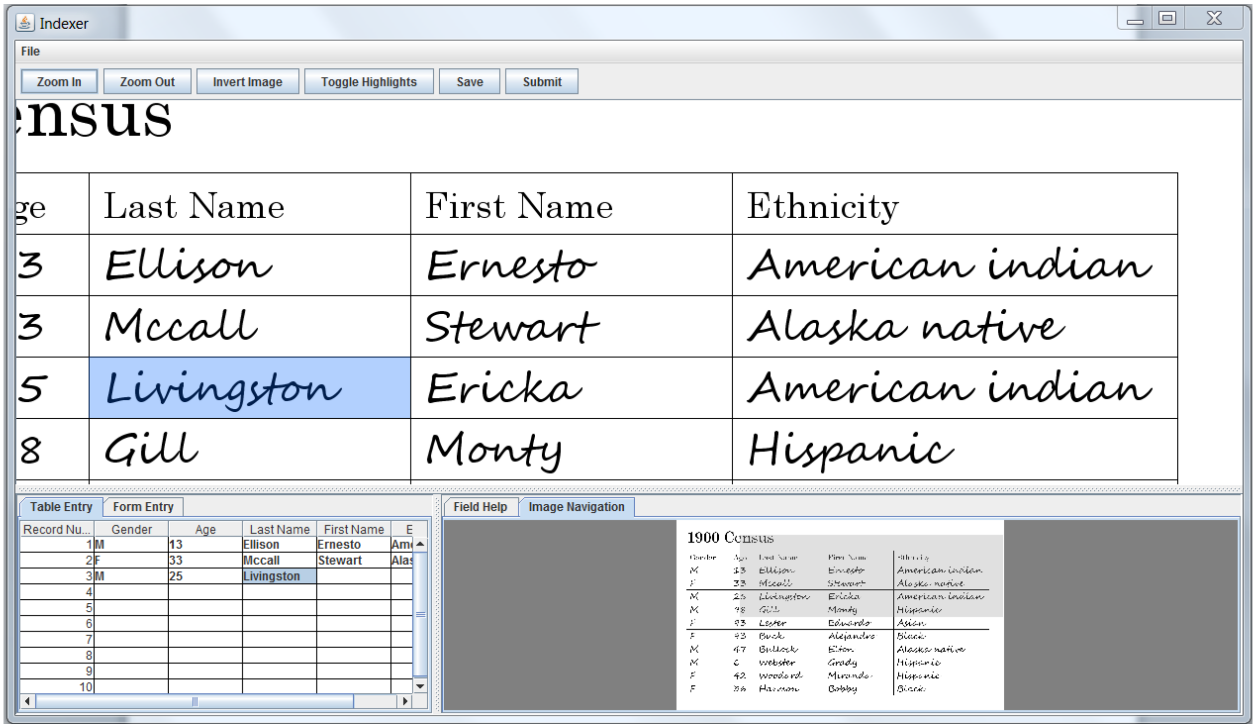
Record Nu...	Gender	Age	Last Name	First Name	E
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Figure 8 - Downloaded Batch

Data Entry

Once they've downloaded a batch, the user can start indexing. The goal is simple: read the words in the batch image and type them in.

The indexing window is divided into three parts (Figure 9). The "image panel", which occupies the top half of the window, displays the batch image that is being indexed. The user can pan (or scroll) the image by clicking and dragging it with their mouse. They can zoom in and out using the scroll-wheel on their mouse or the "Zoom In" and "Zoom Out" buttons. To keep track of their current location on the image, the user can select a "cell" in the image by clicking on it with their mouse. The currently-selected cell is highlighted blue. The blue highlights can be turned off and on by clicking the "Toggle Highlights" button.



ensus

ge	Last Name	First Name	Ethnicity
3	Ellison	Ernesto	American indian
3	Mccall	Stewart	Alaska native
5	Livingston	Ericka	American indian
8	Gill	Monty	Hispanic

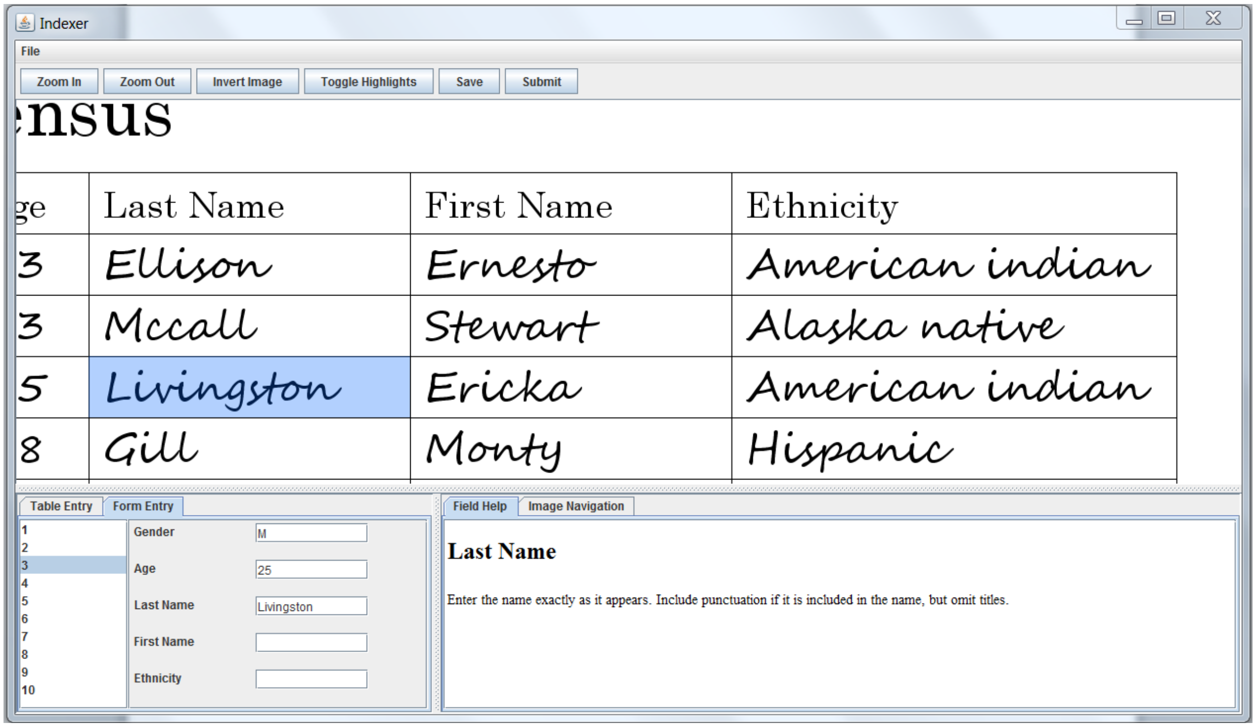
Record Nu...	Gender	Age	Last Name	First Name	E
1	M	13	Ellison	Ernesto	Am
2	F	33	Mccall	Stewart	Alas
3	M	25	Livingston		
4					
5					
6					
7					
8					
9					
10					

Figure 9 – Indexing Window: Image Panel, Data Entry Panel, Help Panel

The bottom half of the window is divided into two parts, the "data entry panel" on the left, and the "help panel" on the right. The data entry panel displays a table where the user can type in the values that they read off the image. The table cell corresponding to the currently-selected word in the image is highlighted blue (just like the word itself is highlighted blue in the image). This helps the user keep track of their current position in the table. The user can move from one table cell to the next by pressing the TAB key. They can also move to any cell in the table by clicking their mouse in the desired cell. When they move from one cell to another, the blue highlights in the table and the image both move to indicate the newly-selected word (i.e., the highlights in the table and image are always in synch).

The help panel (bottom-right) displays the "image navigator". The image navigator displays a birds-eye-view of the image, and draws a gray rectangle over the part of the image that is currently visible in the image panel. This gives the user some context about which part of the image they are looking at in the image panel when only part of it is visible. By dragging the gray rectangle in the image navigator with their mouse, the user can pan (or scroll) the image in the image panel. This allows the user to precisely select the part of the image they want to look at.

The data entry panel also provides the user with the option of typing the data into a form rather than a table (Figure 10). By selecting the "Form Entry" tab, the user can switch from a table view to a form view. In the form view, the user can select a particular record by clicking on the appropriate record number on the left. After selecting a record, they can enter the values for the record in the form on the right. As with the table view, the currently-selected field in the form is kept in synch with the currently-selected word in the image panel, and the user can move from one field to another either by pressing the TAB key or by using their mouse to select the desired field. Users may switch back and forth at will between the table view and form view.



ge	Last Name	First Name	Ethnicity
3	Ellison	Ernesto	American indian
3	Mccall	Stewart	Alaska native
5	Livingston	Ericka	American indian
8	Gill	Monty	Hispanic

Gender	M
Age	25
Last Name	Livingston
First Name	
Ethnicity	

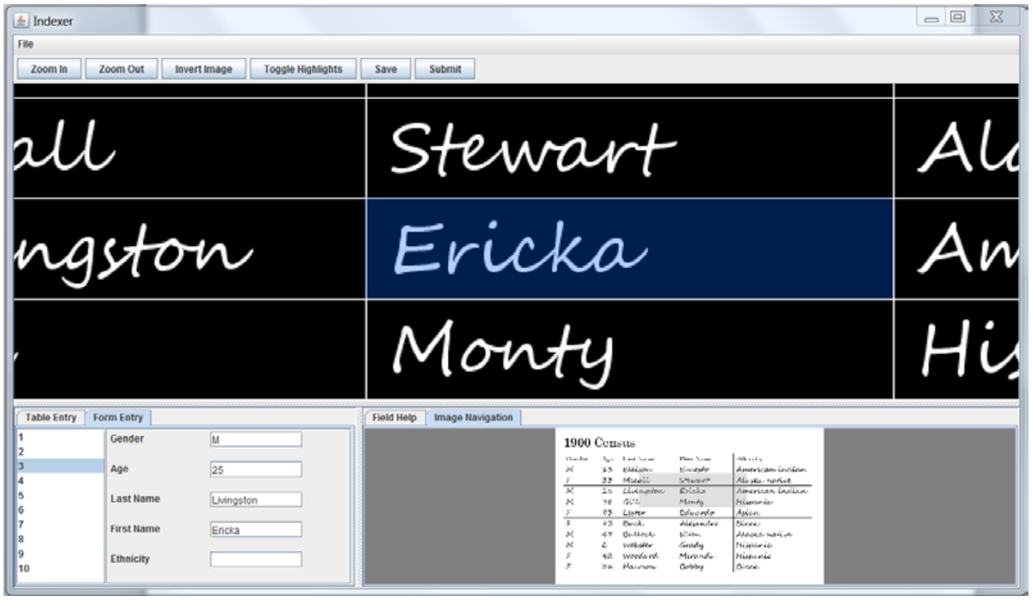
Last Name

Enter the name exactly as it appears. Include punctuation if it is included in the name, but omit titles.

Figure 10 - Form Entry & Field Help

The help panel (bottom-right) also contains two tabs. In addition to the "Image Navigation" tab, it also provides a "Field Help" tab, which displays end-user help about the field that is currently-selected in the image. This field help gives users precise instructions on how to enter values for that field. For example, the help for the "Last Name" field might instruct users to include punctuation but omit titles. Users may switch back and forth at will between the field help and image navigation tabs.

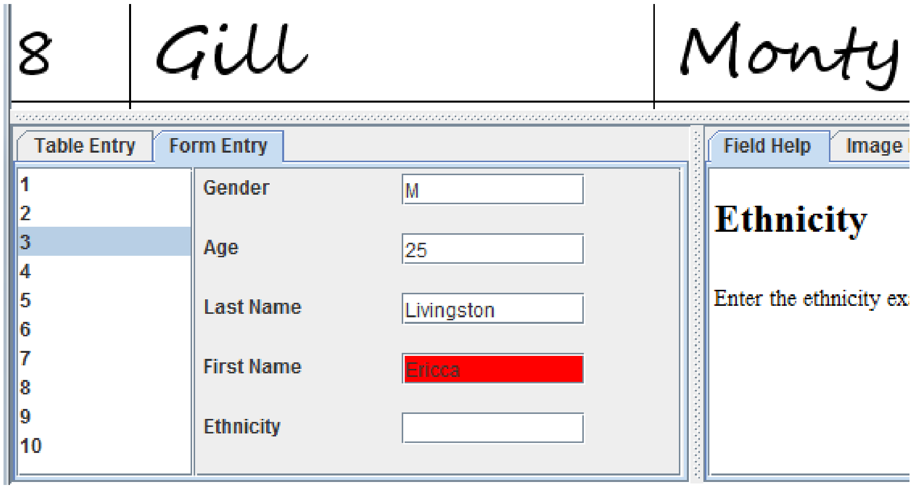
If the user is having difficulty deciphering the handwriting on the image, it can be helpful to "invert" the image. Inverting the image turns all black pixels to white, and vice versa (Figure 11). Sometimes looking at both the inverted and non-inverted images can make it easier to decide what value a field contains.



Gender	M
Age	25
Last Name	Livingston
First Name	Ericka
Ethnicity	

Figure 11 - Inverted Image

The Client contains a "quality checker" that assesses the accuracy of the values entered by the user. When the user enters a value for a field, the Client checks the value against a set of known values for that field. For example, user-provided values in the "Last Name" field in the "1900 Census" would be checked against a set of common last names from that place and time period. If the user-provided value is not found in the set, the Client highlights the value by coloring it red (Figure 12). When the quality checker highlights a value red, that doesn't mean that the user is necessarily wrong. It is only a reminder to check again to make sure that the value is correct, because the system doesn't recognize the value that was entered.



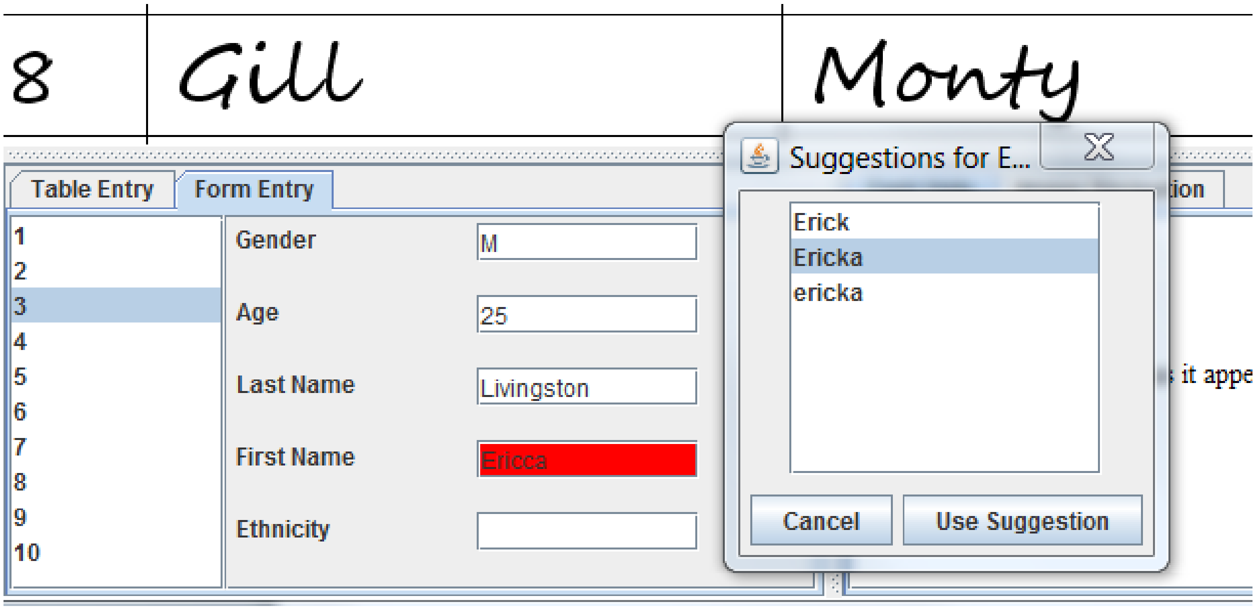
Gender	M
Age	25
Last Name	Livingston
First Name	Ericca
Ethnicity	

Ethnicity

Enter the ethnicity ex

Figure 12 - Quality Checker

By right-clicking their mouse on a red-highlighted field and selecting “See Suggestions”, the Client will offer suggestions of known values for the field that are similar in spelling to the value entered by the user (Figure 13). This can give the user ideas about what value the field might actually contain. (For example, the user might think a field contains the value “Ericca”, and the system might suggest that “Ericka” is more likely.) By selecting a value in the list of suggested values and clicking “Use Suggestion”, the Client will automatically replace the user-specified value with the selected suggested value.



Gender: M
Age: 25
Last Name: Livingston
First Name: Ericca
Ethnicity:

Figure 13 - Suggested Known Values

Save / Logout / Exit

The user can save the current state of their work by clicking the “Save” button at the top of the indexing window. This results in everything the user is doing to be saved locally, including the batch they are currently working on, all of the field values they have typed in, the current zoom and pan settings of the batch image, the size and position of the indexing window on the desktop, etc. The current state of the user’s work is also saved any time they logout or exit the program. Later, when the user starts the Client to resume indexing, they are taken back to exactly the same place they left off the last time they left.

Clicking the “Save” button saves the state of the user’s work, but leaves them logged in.

Selecting the “Logout” option on the File menu saves the state of the user’s work, logs them out, but leaves the Client running. In this case, the indexing window is closed, and the login dialog is displayed so another user can login (Figure 1).

Selecting the “Exit” option on the File menu saves the state of the user’s work, and exits the program.

Submit

When the user finishes indexing a batch, they click the “Submit” button at the top of the indexing window. This causes all of the field values entered by the user to be uploaded to the Server, which stores the values in a database so they can be used in key word searches in the future. The Server also increments the number of records that have been indexed by the user so they can see that their hard work is paying off the next time they login. After the user submits their batch, the Client returns to an empty indexing window, and waits for the user to download another batch (or logout/exit).

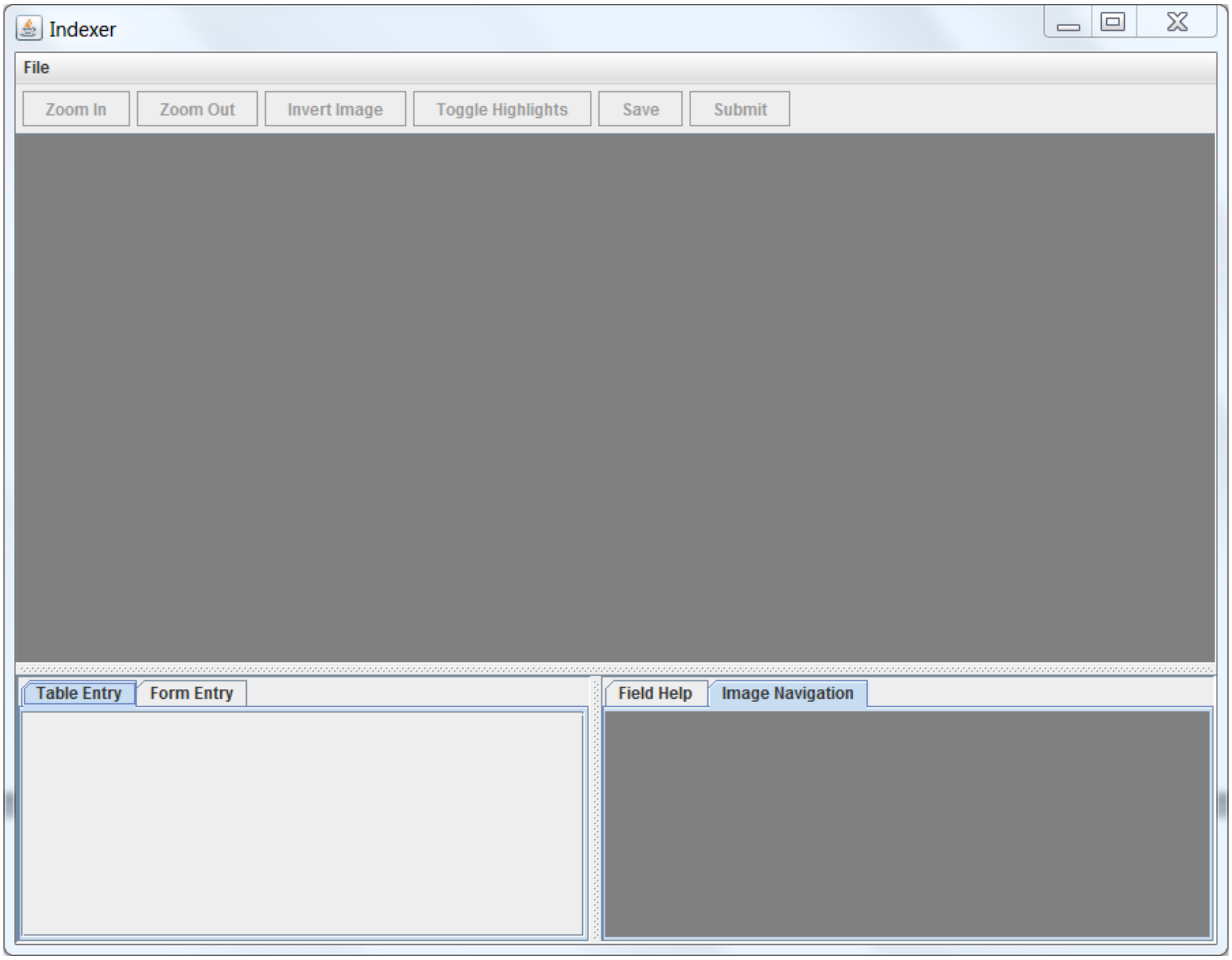


Figure 14 - After Submitting a Batch