The New Immigrant Ancestors Project Software: A Web-Based Approach to Data Extraction and Retrieval

Mark Witmer

February 20, 2009

Abstract

The Immigrant Ancestors Project extracts European Emigration records in many languages and formats. This paper explains a purely web-based solution to extracting the data into a database and coordinating the work of a group of volunteers in several different countries. It details the technology used in the software and the procedure that the project supervisors now follow, beginning with scanned images and ending with a searchable online database.

1 Introduction

1.1 About the Immigrant Ancestors Project

The Immigrant Ancestors Project (IAP) is a research project sponsored by the Center for Family History and Genealogy at BYU dedicated to the retrieval, extraction, and digitization of European emigration records in order to make the information available to academic researchers and individuals searching for their ancestors. The project employs a number of student researchers and also coordinates the work of volunteers from around the world, who extract records in their native languages. To date, the project has worked with records in seven languages: English, German, French, Italian, Portuguese, Spanish, and Dutch. A large variety of sources, record formats, and languages created the need for software that was flexible enough to meet the needs of each language section and record collection.

One distinction between the Immigrant Ancestors Project and other similar projects is that the original documents are not readily available to researchers. In most cases, the agreements between the participating archives and the Center for Family History and Genealogy dictate that only the extracted data and not the original images are to be displayed on the IAP website. For this reason,

rather than merely indexing the documents, volunteer extractors perform a full digitization of every relevant field.

Because the original documents cannot be viewed from the public search, the contents of the database are the principal product of the Immigrant Ancestors Project. Owing to the crucial role that this data serves in the project's mission, its accuracy is of the highest importance. For this reason, trained supervisors examine each extracted record individually to verify its correctness before it is made available to the public. While this approach encourages accurate transcription, it inevitably limits the number of volunteers that the project can utilize effectively. Consequently, this approach is better suited for the small collections of detailed emigration records that IAP focuses on than for large and consistently formatted vital records.

1.2 Previous Software

Previous software for the project was created in an *ad hoc* manner, based upon immediate need without regard for creating a single unified application. The result was a number of individual but interdependent programs that each fulfilled one step of the digitization project but relied upon the proper functioning of other programs in order to function themselves. Consequently, as evolving requirements led to changes in one or several components of the suite, the system as a whole became unacceptably unreliable.

For example, the image viewer/data entry software that users downloaded onto their computers produced data in a particular xml format. Any changes to that format would have required the modification of the data entry program, the program that loaded the xml data onto the database, and the program that created the files that defined a batch's fields. Each of these programs was written in a different language, by a different programmer, and at a different time. Furthermore, certain necessary steps in the digitization process, such as transferring extracted data to a relational database, never reached a satisfactory state of completion to begin with.

1.3 Requirements for a New System

Recognizing the inadequacy of the current generation of software, the Project leadership decided to produce a new application which would combine the features of the old suite into one application that was to be, as far as it was possible, written for a single platform and language. Following is a brief summary of the features required in this system:

- A user management module giving supervisors the ability to manage the user ids, profiles, and permissions of their volunteers and other supervisors
- A data extraction module where supervisors and volunteers can view record images and transcribe the information they encounter into a number of

previously determined fields unique to each record collection

- A method for supervisors to define the fields that volunteers are to look for in the collections
- A way to translate the text in the application to the different languages spoken by volunteers and visitors
- An advanced search that allows people to search the extracted records by name, event dates and locations, and other attributes
- Documentation detailing the nature and extent of the records available from the Project
- A way for supervisors to examine and verify any field from any record in the database, and remove data from the public search until corrected if errors are discovered.

The rest of this paper describes the design of the new software system and how it meets the project's needs.

2 Application Design

2.1 A Web-Based Approach

Much of the work in the Project requires contributions from individuals spread across different nations and even continents. As a result, some use of the Internet for the transfer of data and images is without question necessary. However, the extent of its use as a software platform was somewhat limited under the previous system. The extraction software for volunteers was a program written for Windows that they downloaded and ran on their systems; it doubled as an image viewer and as a web client for downloading images and submitting completed batches. This approach limited compatibility to Windows; in fact, the introduction of Windows Vista led to compatibility issues that caused serious complications for some volunteers.

Based on experience with the previous extraction client, the new software was designed to make sole use of the Internet as its platform. The application runs on a centralized server and volunteers and supervisors alike interact with it through their web browser. The result of this decision is that any individual with an Internet connection and a reasonably modern web browser such as Internet Explorer 6+, Firefox 2+, Opera 9, or Apple Safari 3 can participate in the project as a volunteer.

2.1.1 Ruby on Rails

The project's new software was written in Ruby on Rails, a web development framework for the Ruby programming language. The advantages of that framework for this project include its limited structural overhead, iterative database

migrations that allow changes to the database schema throughout the development project, and the ability it gives to quickly implement an idea.[1] IAP only employed one web developer during the time of this project, so it was essential that the web framework be lightweight and small, so that one person could develop the entire application. Because the application itself was not of such a great size, a larger development framework, such as Microsoft's .NET or Enterprise Java, would not have been needed or wanted. Other websites that have followed a similar approach include Twitter,[2] Hulu.com, and Geni, a genealogy-based social networking site.[3]

Rails applications refer to dynamically generated web pages as *views* and groups related views together in a *controller*. Below are discussions of several of the controllers involved in IAP software.

2.2 The Archive and Collection Controller

The process of taking scanned or photographed images from an archive and producing digital records of the information that they contain involves several steps. First, the supervisor must specify what archive the images were scanned from, and then what record collection in that archive they belong to. While the Immigrant Ancestor Project does not offer the original images for public download, each record must specify its exact source (catalog number, page number, etc.) so that interested parties may contact the archive and request copies of the original document. Source information is stored for each collection, each batch within the collection, and also in the individual records.

Figure 1 illustrates the archive management view. Each archive has a name and a description that visitors will see when they view a search result that belongs to the archive. Supervisors may also create a form letter that researchers can send to the archive to request images.
The collection management view (Fig. 2) is similar to the archive management view; supervisors specify what collections are being extracted from the selected archive.

2.3 The Batch Controller

Once the supervisor has created a collection in the system, he or she separates related images from that collection into groups called batches. Each batch can be assigned to a volunteer for extraction.

The batch controller (Fig 3) displays which batches are newly created, which have been assigned but are not yet extracted, which are extracted and awaiting verification (examination by the supervisor for possible corrections), and which

[1] Lenz, Patrick. (2008) Simply Rails 2 Collingwood, VIC, Australia: SitePoint Pty. Ltd.
[2] "How is Twitter Built", http://twitter.com/about
[3] http://rubyonrails.org/applications

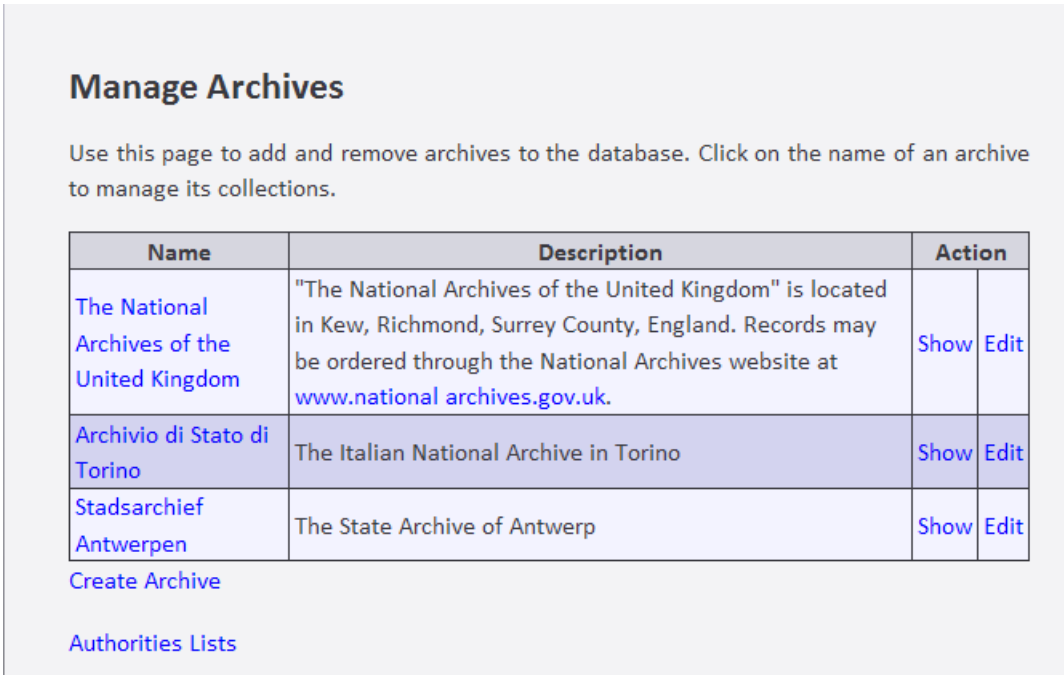

Figure 1: Managing Archives

are verified and thus visible in the system's public search. The user can click "view" in order to view more information and options for a specific batch. Each collection has a model batch, where the supervisor specifies the expected fields to extract and of which all newly created batches are copies.

Batches from the previous software are converted into an XML file and imported using the "Import XML" tool. When the current batches are fully integrated into the new software, this tool will no longer be necessary, as all extracted data is immediately on the database, and simply needs to be marked as verified for the public search to refer to it.

Additionally, the controller shows a list of images which are not assigned to a batch (in the example below, all images are assigned to a batch). Finally, each batch can belong to a batch group, which specifies more source information about the records in that batch.

2.4 The Extraction Controller

The extraction controller consists of an area to select, view, and zoom in and out of images, a form to enter information from the images, and a menu at the top that shows different options depending on the permissions of the user.

The data entry form has two similar implementations; one, the record view (Fig. 4), displays extraction fields for one person at a time. This view allows users to see a maximum of information about a single individual without the need to

Figure 2: Managing Collections

scroll excessively. Additionally, as shown here, when the user is an authorized supervisor, he or she has the option to add or remove fields from the record. Initially, the supervisor will add the expected fields to a collection's model batch and records created thereafter will simply copy the model batch. Subsequently, the supervisor may edit other batches and manually add or remove fields that are or are not required in individual cases. In this manner, each record in the database can represent a unique set of attributes, rather than a solely uniform group of predetermined fields. The other option is the table view (Fig. 5), which displays multiple records in tables beneath the image. If sequential records do not have matching fields, a new table begins with the expected fields for that table. This view is more helpful than the record view during the actual extraction and verification section, as volunteers are able to add or remove records without having to constantly select which person to view and load the associated information. Users with a screen resolution of 1024x768 pixels or greater can use either view comfortably; at 800x600 the software is still usable, but the size of the extraction form limits the portion of the image that can be viewed at once.

One feature often found in extraction software that is not present in the IAP software is automated field highlighting. The large number of documents that are handwritten letters lacking a consistent format make this feature impractical: the software would highlight a value other than the one desired so often it would generate more confusion than clarity.

Data extracted by volunteers requires verification to ensure its accuracy. Sev-

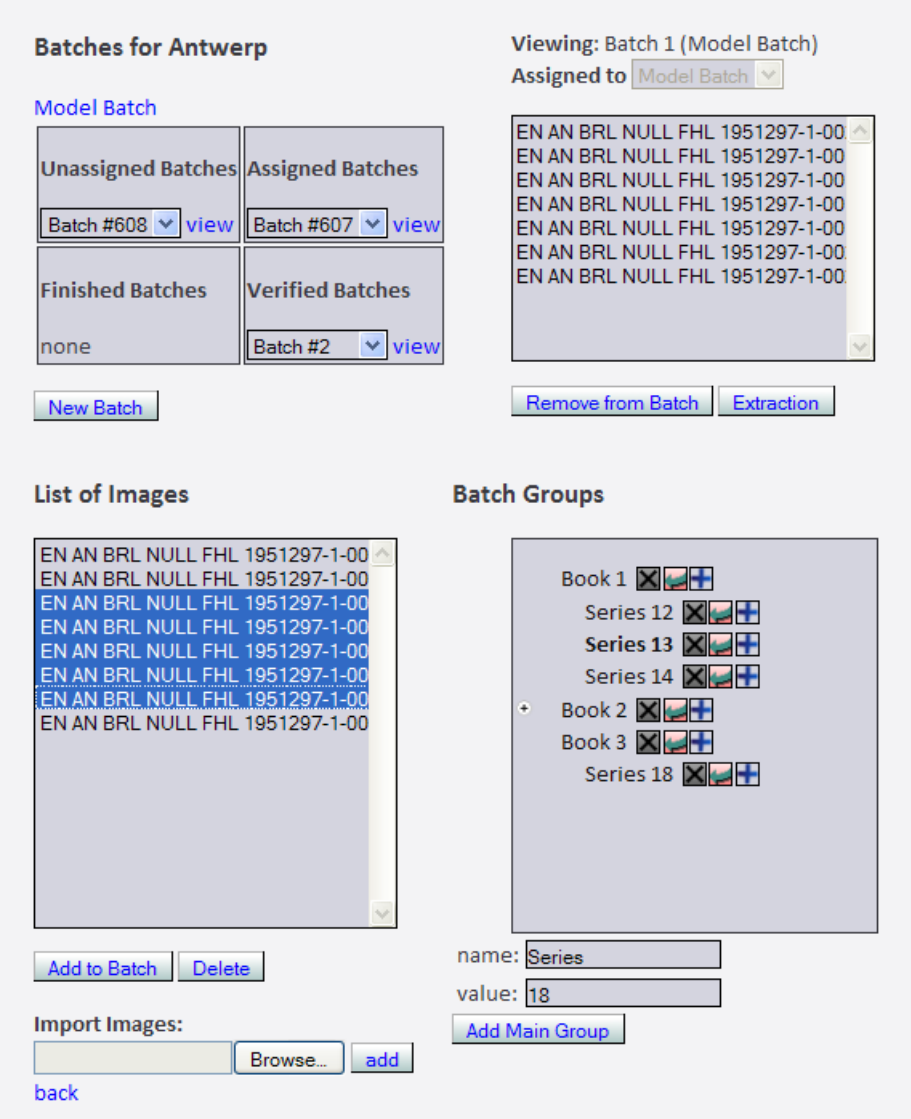

Figure 3: Managing Batches

eral methods exist for the verification of extracted data, including its examination by a second individual, or extraction by two people and correction of any discrepancies in the results by a third party. Due to the complexity of many of the documents in the project, verification is performed by trained supervisors familiar with the documents in use. They reopen the extraction controller for a finished batch, check that the extracted information is correct, and then mark it as verified. At that point visitors to the project web site can search for the extracted information.

These views are only available to supervisors and volunteers: in particular,

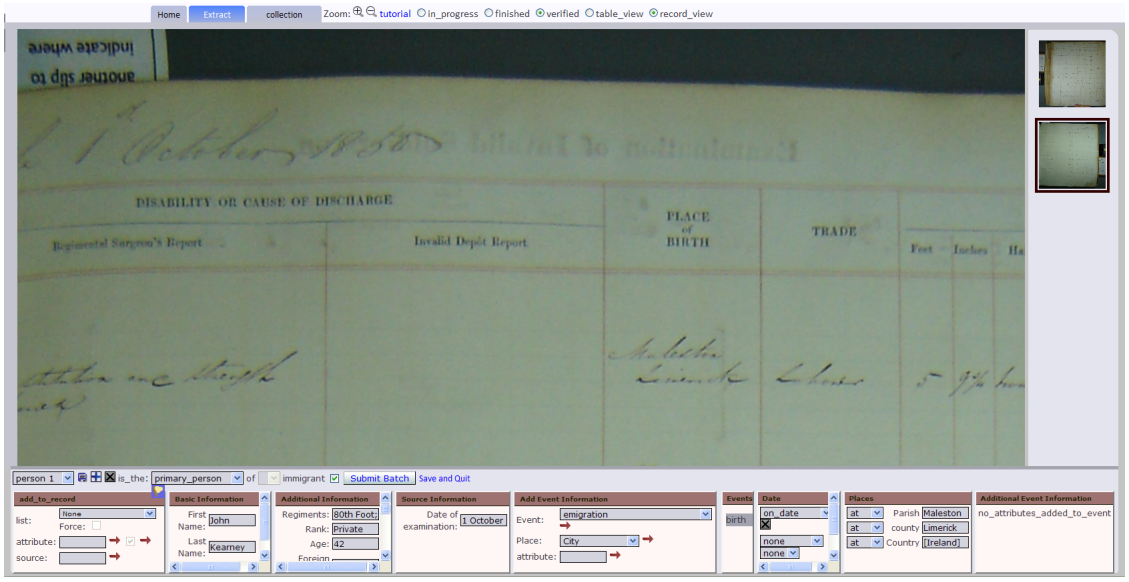

Figure 4: The extraction record view

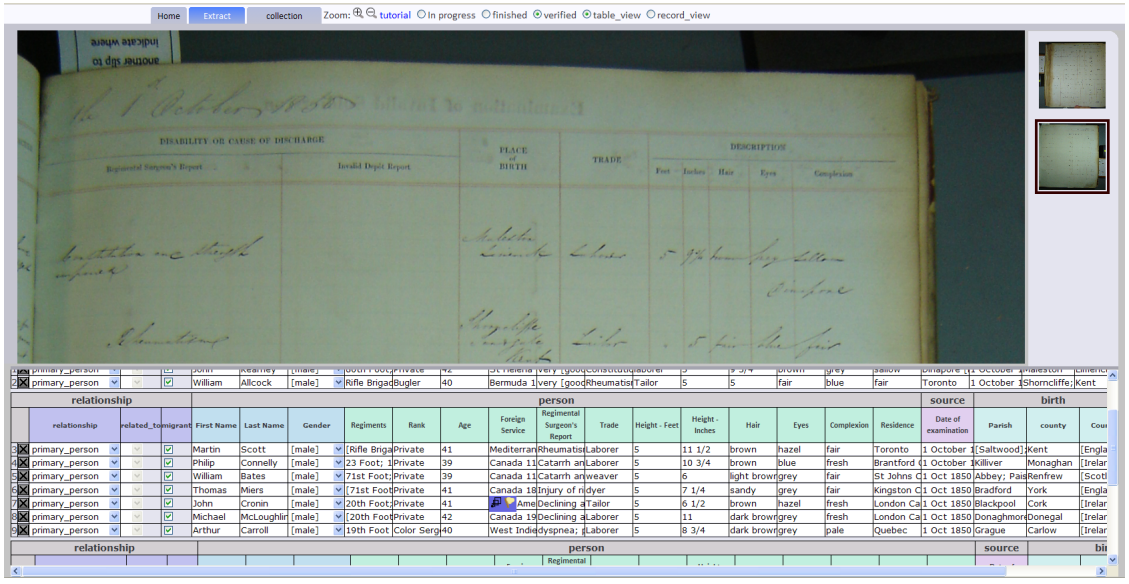

Figure 5: The extraction table view

the images themselves cannot be accessed without proper authentication. They do not have a public URL because the project's agreement with most archives prevents the public distribution of scanned images.

2.5 The Users Controller

New volunteers can sign up with IAP from a link on the front page. Anyone can volunteer, although before they are permitted to proceed with extraction at their own pace they must have a supervisor assign and check their first batch to ensure that their skills are sufficient for the documents at hand. Supervisors are able to view and modify the accounts of volunteers in the Users controller (Fig 6), which divides them into sections, each corresponding with one of the languages involved in the project. The view also specifies the last date that a volunteer did any extraction. Volunteers that have been inactive for more than two weeks have the date marked in red to suggest that they should be contacted and/or removed.

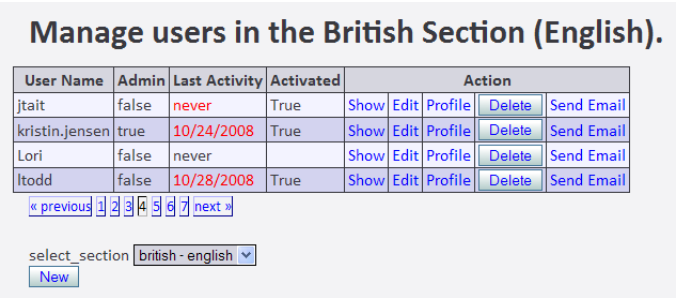

Figure 6: Managing Users

Volunteers also specify some information in a profile that aids supervisors in assigning them to a particular collection (Fig 7).

2.6 The Search Controller

The public interface with the application is the record search (Fig 8). Visitors can specify information about a person's name, specific events of interest, and dates and places for those events, and the search will return a list of records that match the provided parameters (Fig 9). If a user clicks on a particular person in the results, a view comes up with all the available details about that person (Fig 10). Merging different records of the same person is out of the scope of this project, so each record represents a unique reference to a person in a particular document.

Each collection also contains a small group of sample images that give users an idea of what sort of document they would receive if they were to request the original document from the archive. Users can view these images by clicking "View Sample Images" by the collection name. Some records also allow users to create letters to send to the archive in order to request the originals.

Figure 7: User Profile

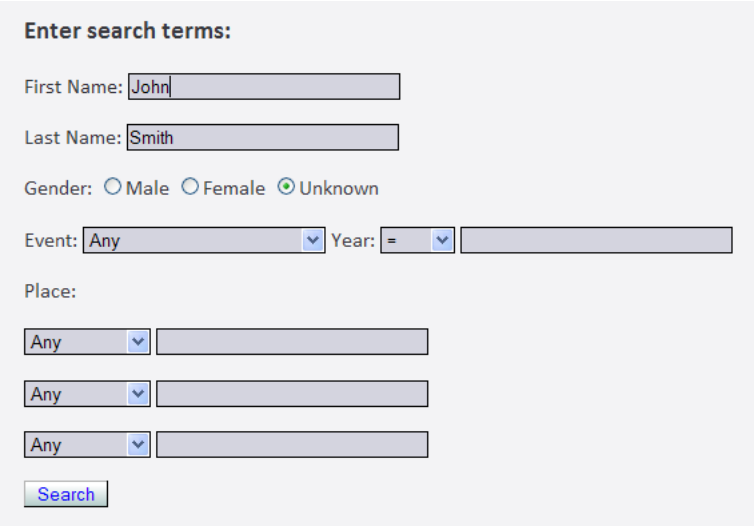

Figure 8: Search Terms

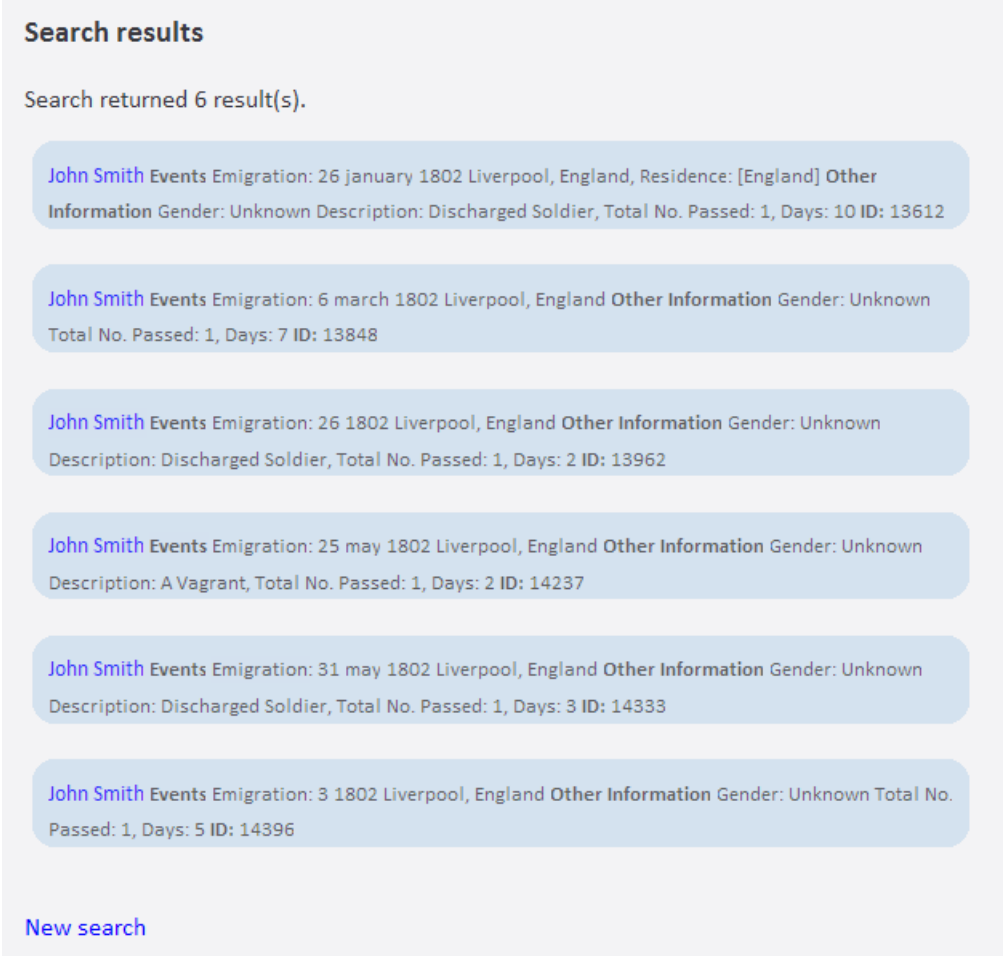

Figure 9: Search Results

2.7 The Forum Controller

Users can post questions and comments about the software and their assigned batches in the forum section of the software (Fig. 11). Both other volunteers and supervisors can respond to these posts in order to provide rapid feedback to users and keep a record of previously resolved issues. This record will assist new volunteers in answering common questions on their own by browsing previous posts in the forum.

Joaquim Fernandes Querido	
Gender:	Male
Emigration:	True
Age:	31 anos
Occupation:	Trabalhador
Marital Status:	Solteiro
Physical Description:	True
Validity of Passport:	sessenta dias
Events	
Passport Issued:	14 November 1871
Birth:	
at:	
county:	Vila Nova de Poiares
country:	[Portugal]
Born in:	Arrifana
District:	[Coimbra]
Emigration:	
to:	
city:	Santos
state:	[São Paulo]
country:	[Brasil]
Port of Departure:	Lisboa
Source	
archive:	Portuguese Archive
An archive of Portuguese records	
collection:	Coimbra - View sample images
A collection of records from Coimbra	
Record Number:	873

New search

Figure 10: Record View

3 Other Features

3.1 Internationalization

As previously mentioned, the Immigrant Ancestors Project has records in seven different languages. In order to accommodate visitors in all of those languages,

Figure 11: The Forum Controller

the system makes use of an in-place translation utility that allows supervisors to translate content into any of the languages. Upon first visiting the site, users are invited to select the language they prefer, and from that point all of the content will be translated into that language. Supervisors have the additional option of right-clicking on a piece of text or form item and bringing up a translation dialog box (Fig. 12) where they can change the value for that key in the current language. When the change is finalized, the text on the page will be updated to reflect the new value.

3.2 Static Content

In addition to the web application for sorting, extraction, and searching of emigration records, the IAP site contains several resources describing the nature of European Immigration, other available resources, and information about its legal and social causes and consequences. All of this content is available from the same page as the searchable database.

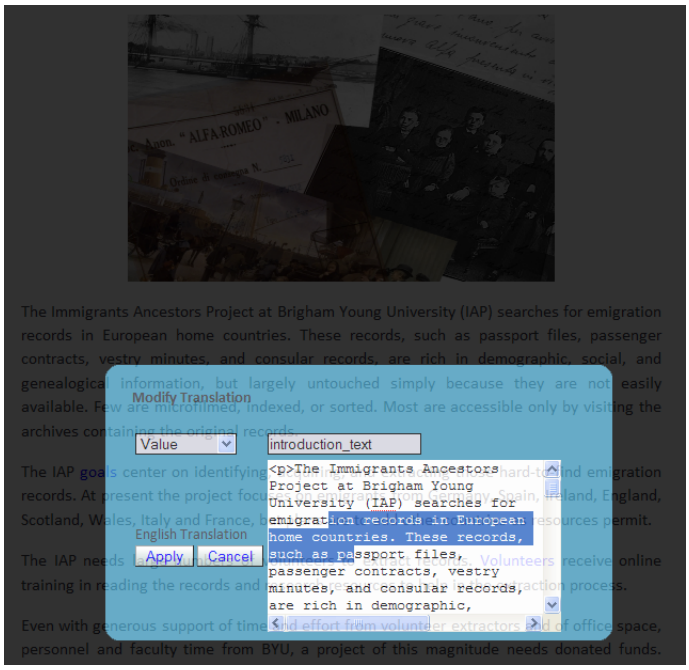

Figure 12: Translation View

3.3 Future Additions

3.3.1 Maps

As the database of records increases in size, the project will provide a mapping tool where users can specify event types, dates, and places, and then see how many people traveled to different places within the parameters specified. This tool will aid researchers who are looking for large-scale trends in migration patterns, and will give them an empirical tool to sift through the large amount of data that will be available.

3.3.2 Widgets

Another feature under consideration is a freely distributed widget for individuals interested in making the IAP search available to visitors to their web sites. They would be able to insert some code into their sites that would give visitors a search form which would display results and refer them to the IAP site to view the records.

3.3.3 Advanced Image Viewer

In order to assure compatibility with as many volunteers' computers as possible, the image viewer is implemented in a javascript file that runs inside their web

browsers. This approach prevents the viewer from having features such as rotation, brightness, and contrast adjustment, and other image manipulation tools that are often useful. Later versions of the software should include an alternative viewer that makes use of a rich application client such as Flash or a Java applet.

4 Conclusion

The new software for the Immigrant Ancestors Project represents a significant step forward in its embrace of new technology, accessibility, ease of use, and low maintenance requirements. It will hopefully be the vehicle whereby millions of emigration records become available to the general public, helping advance research on the topic for academic researchers and family historians alike.

Appendix

A Comparison With Other Extraction Programs

Below is a table comparing some features of the IAP software, the Church of Jesus Christ of Latter-day Saints's FamilySearch Indexing program, and The Generation Network's World Archive Project.

	IAP	**FSI**[4]	**WAP**[5]
Verification	1 Volunteer extracts, 1 paid supervisor verifies	2 Volunteers extract, 1 volunteer arbitrates	2 Volunteers extract, 1 volunteer arbitrates
Communications with Volunteers	Two-way via email and online forum	One-way from project leadership to volunteers	Online forums
Software Platform	Web-based, cross-platform	Downloaded java applet	Windows Only, support for others planned
Scope of Extraction	Full	Index	Index
Project Size	Several part-time employees, 30-40 volunteers	Several full-time employees, thousands of volunteers	Currently in beta, growing

B Site URL

At the moment, the current IAP site (http://immigrants.byu.edu) continues to run on the old software. The new software is undergoing final beta testing and

[5] Family Record Extraction Administrative Handbook, Salt Lake City, Church of Jesus Christ of Latter-day Saints, http://fch.ldschurch.org/WWSupport/Documents/ FamilyRecordExtractionAdminHandbook_30985_000_000.pdf, accessed 19 Feb 2009

[5] "About the Ancestry.com World Archives Project", http://landing.ancestry.com/wap/learnmore.aspx, accessed 20 Feb 2009

can be found at http://new.iap.byu.edu. Its full roll-out is anticipated in the Spring of 2009.