J Forensic Sci, January 2008, Vol. 53, No. 1
doi: 10.1111/j.1556-4029.2007.00612.x
Available online at: www.blackwell-synergy.com

Lori E. Baker,[1] Ph.D. and Erich J. Baker,[2] Ph.D.

Reuniting Families: An Online Database to Aid in the Identification of Undocumented Immigrant Remains*

ABSTRACT: The Reuniting Families project attempts to aid federal, state and local agencies currently working towards the identification of deceased undocumented immigrants. This initiative has created a distributed on-line database, accessible by public officials and private citizens interested in searching for missing individuals based on both phenotypic and genotypic characteristics. This broad effort includes the exhumation of individuals from geographically disparate pauper graves, the classification of their physical characteristics, and the cataloging of observed metric traits in a local relational database, to include associated articles of possession and related metadata. Concurrent with the documentation of physical forensic evidence is the analysis of mitochondrial DNA sequences. Computational techniques and scoring parameters are applied to automate the process of discovery and identification as well at to preserve information on the missing. The result is a prototype knowledgebase that may serve as a model for future efforts in international forensic science collaborations.

KEYWORDS: forensic science, forensic anthropology, mitochondrial DNA, database, undocumented immigrants, immigrant death

The purpose of this project is to provide a common informatics infrastructure designed to assist in the identification of deceased undocumented immigrants by bringing together vastly dispersed families and agencies. This effort is necessitated by a multitude of logistical issues plaguing the forensic science and law enforcement communities.

Every year more than a million individuals pass illegally between the United States and Mexico (1). Legal and illegal immigration has increased dramatically despite an anticipated decrease associated with the North American Free Trade Agreement (NAFTA). These trends may continue with the implementation of the Central American Free Trade Agreement (CAFTA). While illegal immigrants come from a host of countries, Mexican nationals comprise the majority; the number of Mexicans residing in the U.S. was 11.2 million as of March 2004 and 47%, or 5.2 million, of those immigrants were undocumented (2).

The implementation of several enforcement operations along the Southwest border (Operation Hold the Line in El Paso, Operation Gatekeeper in the San Diego area, Operation Safeguard in the Tucson area and Operation Rio Grande in the South Rio Grande Valley) have all served to stop the most actively traversed areas of illegal border crossing (3). In addition, to guard against future terror threats the U.S. Customs and Border Protection Agency became part of the newly created Department of Homeland Security in 2003. Border security has been augmented along the roughly 2,000 mile U.S./Mexico border as a result of this increased federal priority. Consequently, there remain few urban migratory routes not well patrolled and these are confined to areas of extremely inhospitable terrain. Migrants that travel these more secluded routes parish at an alarming rate. For example, during the first 8 months of 2005, there have been 358 reported immigrant deaths (4).

Unfortunately, there is no consensus among agencies of the total number of border deaths or how many of those individuals remain unidentified. While many deaths undoubtedly are undiscovered due to their remote locations, inefficient recording and reporting may cause the bulk of inaccurate statistics. For example, the U.S. Border Patrol began tracking migrant deaths in mid-1998 with the creation of the Border Safety Initiative. A total of 1,733 migrant deaths were reported by the Border Patrol up to July, 2003, and 769 of these were reported as unidentified, roughly 44%. The number of deaths reported here may be vastly undercounted, in part, because reporting deaths to border patrol officials is not compulsory (5). With no single accountable agency, the tabulation of illegal immigrant deaths is problematic.

Indeed, the authors believe the major contributing factor to the difficulty of solving forensic cases of unidentified individuals is the lack of standardized data reporting and storage among agencies. To date there is no centralized, private, federal or state-mandated mechanism that attempts the collection, curation or repatriation of the undocumented immigrant remains as a unique group of individuals. However, there are many different agencies working independently and in *ad hoc* collaborations to identify and return the deceased to their families. Ultimately they are hampered by inefficient data sharing and accessibility to digital infrastructure. In most cases at the local level, a lack of resources precludes modern informatics approaches. Inquiries require the physical location of files which are inherently difficult to search, a painstakingly slow process for counties that have a large number of deaths.

Lastly, funding for forensic analysis is limited for local and county agencies. Due to cost constraints, numerous remains are never analyzed by medical examiners or forensic experts. DNA typing is also cost prohibitive. Some counties retain bone samples for potential future testing but this is not standard procedure.

[1] Department of Anthropology, Forensic Science and Archaeology, Baylor University, Waco, TX 76798.

[2] Department of Computer Science, Baylor University, Waco, TX 76798.

Received 30 Dec. 2005; and in revised form 5 Nov. 2006; accepted 1 July 2007.

Generalized Reuniting Families Database Schema

The Reuniting Families (RF) database (http://www.reunitingfamilies.org) leverages existing open-source informatics technologies to collect and collate physical, phenotypic, and genotypic information from geographically disparate agencies. The overall goal of the system design is to allow clients and administrators the ability to search controlled data repositories in a two-tier format: (i) general matches to physical evidence; and (ii) specific identification based on molecular criteria. This two-tier approach is necessary because it is cost prohibitive to perform molecular tests on the large volume of public queries handled by RF.

Different cohorts of users require differing levels of interaction with the underlying data. For this reason, the database offers multiple interfaces. The public user interfaces allow users to search unidentified remains based on a physical description of the missing individual. Here, they have access to certain images of artifacts but do not have access to sensitive images or controlled information. Administrators and Contributors to the RF database (medical examiners, forensic anthropologists) have broad abilities to upload information and perform advanced searches of the data space. The overall schema, as seen in Fig. 1, represents the relationship of different users, data interoperability, and types of data collected.

From the perspective of a public user, the flow through the Reuniting Families data system is described in Fig. 2. Briefly, after a new user is registered, they may submit a detailed web-based form describing contact information, physical characteristics, secondary articles, including clothing and jewelry, and pertinent free-form comments relating to the missing individual. If the query engine discovers an appropriate match in the database of physical remains, the report of a possible hit is forwarded to a database administrator. A secondary examination of the same search query by the administrator will determine if further DNA testing is warranted. If the automated search does not produce a possible match, the query may be stored for searching at a later time, either through a user-initiated or automated process.

In the event of a DNA examination, RF-associated modules are designed to test, store, and analyze the mitochondrial DNA (mtDNA) sequences from Hypervariable Regions I and II (HVI and HVII). Maternally related family members submit reference samples that are subsequently compared to the unidentified individual. Matching mitochondrial sequences plus strong circumstantial evidence are indicative of identification, resulting in notification of the family, the submitting contributory user, and repatriation of the remains. Reference sequences from nonmatches are stored in the system and compared to each successive sequence from unknown individuals.

Database Features

In addition to web-based interfaces for data upload, searchers, and automated reporting, the Reuniting Families database system contains several tools that greatly augment its utility.

Stored Queries

Once a user performs a query of the database for possible matches, that query becomes a persistent object attached to the specific user account. This allows the user to modify the query for future user-generated searches. In addition, the RF database periodically performs mass queries of all stored searches. This is designed to take into account the inclusion of additional data.

Multi-level User Authentication

Because the Reuniting Families data system is implemented as both a central data collection warehouse for sensitive information related to human remains and a public multi-user query platform,

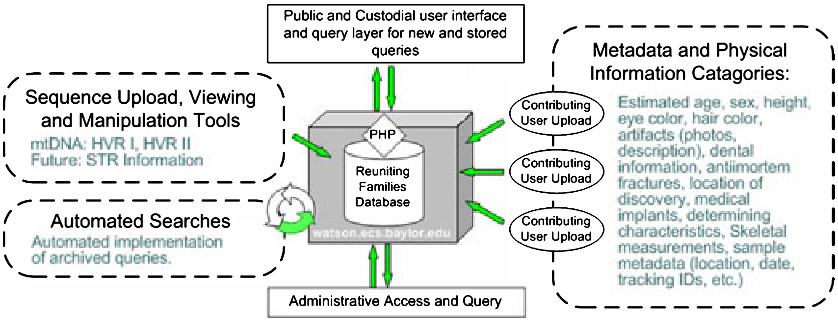
Public and Custodial user interface and query layer for new and stored queries
Sequence Upload, Viewing and Manipulation Tools
mtDNA: HVR I, HVR II
Future: STR Information
Automated Searches
Automated implementation of archived queries.
PHP
Reuniting Families Database
watson.ecs.baylor.edu
Contributing User Upload
Contributing User Upload
Contributing User Upload
Metadata and Physical Information Catagories:
Estimated age, sex, height, eye color, hair color, artifacts (photos, description), dental information, antiimortem fractures, location of discovery, medical implants, determining characteristics, Skeletal measurements, sample metadata (location, date, tracking IDs, etc.)
Administrative Access and Query

FIG. 1—*Overall schema of the Reuniting Families database. The Reuniting Families system is based on an open-source data model, including dynamically created database-backed web pages. User interfaces are proved for Public and Custodial accounts, and upload services allow Contributing users to supply metadata and physical descriptions of unidentified remains. Molecular analysis and viewing tools are available for Administrators, and automated periodic queries of stored searchers ensure that newly added database information is examined.*

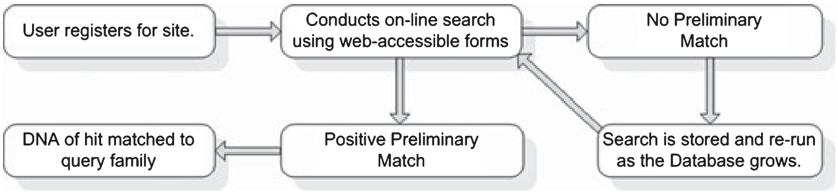
User registers for site.
Conducts on-line search using web-accessible forms
No Preliminary Match
DNA of hit matched to query family
Positive Preliminary Match
Search is stored and re-run as the Database grows.

FIG. 2—*Generalized schematic of public user flow. New users are registered and may submit information about missing individuals. If the query engine discovers an appropriate match to physical remains, the report of a possible hit is forwarded to a database administrator. A secondary examination of the same search query by the administrator will determine if further DNA testing is warranted. If the automated search does not produce a primary match, the query may be stored for searching at a later time, either through a user-initiated or automated process.*

access is controlled through a hierarchical authentication system. As a result, the dynamically backed database structure supports a four-tier approach to data accessibility and integration. Users are classified as (i) administrators, (ii) contributors, (iii) custodians, and (iv) public.

i. Administrators: Administrators have complete access to Reuniting Families database content. They create and delete user accounts for other Administrators, Contributors, and Custodians, perform advanced command-line queries, delete or alter incorrect or redundant database information, and view raw data, including molecular information, database statistics and user load.
ii. Contributors: This category includes individuals from medical examiners offices, forensic anthropologists, and other designated information procurers. Contributors have the ability to perform on-line searches using web-based forms and upload content using pre-formatted HTML forms. Contributors cannot remove or alter content once it is uploaded.
iii. Custodians: These accounts represent officials from law enforcement or nongovernmental organizations (NGO) that require the ability to search for multiple individuals. Custodians have the ability to search on-line forms for multiple targeted individuals on behalf of third parties. They are created and controlled by administrators.
iv. Public: Public users may register without the requirement of self-descriptive information. Their profile is stored in the RF database along with their searches. Public users wishing to serve as their own contact points are required, at a minimum, to supply an email address.

Data Uploads and Automated Reporting

In order to collect as much information as possible while simultaneously minimizing the overhead for Administrators and Contributors, the RF database accepts uploaded information directly into the database via the Internet. Because the RF system deals with a limited number of active contributors, web-based forms are currently created and customized on a per-contributor basis. The database is currently being expanded to accept raw data (e.g., specific skeletal measurements) and bulk uploads in the form of Microsoft Excel spreadsheets. In the event of a possible match, contributors of uploaded information are automatically notified and may assist in verifying the veracity of the hit.

Automated Match Discovery

A significant portion of the database's power is its ability to indicate possible connections between user queries and the physical remains described within the system. This tool is supported by a point scoring algorithm that assigns a score to each query variable. The score is the sum of a multiple combination of variables: variable significance (Vs), expectancy of precision (Ep), and percent of variable match (Vm), see Equation 1. The significance of positive null values are ignored due to an overall uncertainty of their precision. The base values for this schema are as follows: Vs is a measure of the significance of the variable for identification purposes combined with the fundamental accuracy of its measurement. These represent the maximum score of the variable and range from 1 to 30, with items like "Sex" given a Vs of 30 and items such as "Age between 20 and 30" given a Vs score of 1. This value is multiplied by the Ep, or the perceived ability of someone to accurately report the information. It is an integer ranging from 0.1 to 1.0. "Sex," for example, has an Ep of 1.0 while height has an Ep of 0.8. Finally, these variables are multiplied by the Vm field, or the percent match. The Vm for "Height in Centimeters" of an individual measuring to 200 cm as compared to a query field of 210 cm is 0.95. Summing all variables gives the Total Score. Currently, a total score of 75 indicates a possible match.

$$\text{Total}_{\text{score}} = \sum_{0..i\text{variables}} (\text{Vs}_i \times \text{Ep}_i \times \text{Vm}_i) \qquad (1)$$

These criteria are subjective, but they are not intended to provide positive identification. They are designed to automatically flag samples that are of enough interest to be examined manually by an administrator, while minimizing the number of false negatives. The creation of a stable algorithm is an ongoing process.

Database Specifications

The relational database, dynamic web page content, algorithms, and site statistics are performed on a Dell PowerEdge 1650 with dual Intel Pentium III processors running at 1.4 GHz, with 512 Cache and 1 GB 133 MHz SDRAM. This machine operates in tandem with an identical RAID5 backup system. Both servers use the Red Hat Enterprise Linux WS operating system, release 3. We have chosen PostgreSQL version 7.3.10-RH (6) as our RDBMS for its durability and flexibility, and use Apache 2.0.52 (7) to serve web pages created with PHP version 5.0.4 (8). Mitochondrial sequence analysis employs the BLAST set of tools (9), and the phenotype matching algorithms are written in a combination of Perl version 5.8.4 (10) and Java (11) with the support of documented BioPerl modules (12).

Results and Discussion

Family members of missing and deceased immigrants are often frustrated in their attempt to obtain information. A majority of families do not have the resources to search the vastly distributed border agencies on their own. They are typically aided by overwhelmed foreign consulate officials and medical examiners. In addition, forensic scientists are hampered in their identification efforts by limited budgets, large numbers of cases and, in some cases, incomplete phenotypic and genetic data sets required to make assertions about heritage. Combined with a lack of institutional integration at the federal, state and local levels, efforts to repatriate unidentified remains to their families are often unsuccessful.

The purpose of this project is to provide a common environment designed to assist in the identification process by bringing together vastly dispersed families and agencies. The password-controlled on-line relational database simultaneously achieves three objectives. First, it provides a mechanism by which case data can be stored, searched and shared electronically by agencies, creating a real-time working collaboration. Second, as the curated repository grows, it will provide a much-needed set of genetic data and phenotypic data that will aid scientists in future identifications, both in known immigrant cases and in other forensic situations. Third, relatives of missing immigrants are able to participate in an on-line search of the database using unique physical characteristics (e.g., height, age, sex, skeletal trauma) and descriptions of articles of clothing to help identify missing individuals. If a search provides a possible match, DNA testing of a maternal relative will be done to confirm the potential identification, and steps toward the repatriation of the remains will be taken. The performance of genetic analysis on all unidentified immigrant remains and their respective storage will allow future identification and, perhaps, even inadvertent matches to close family members.

Since its inception in late 2003, the Reuniting Families project has aided in the identification of four previously unidentified individuals. While direct discoveries made via queries of the RF database have not yet occurred, the number of active Public users continues to grow along with the procurement of searchable physical and molecular data sets. There are currently over 80 active users and over 100 physical cases awaiting inclusion in the database.

Beginning in the spring of 2006, an attempt will be made to increase the amount of RF database information by focusing on the extensive number of pauper graves that are scattered along the border region. Because most municipalities do not catalogue physical or molecular evidence prior to interment, it may be necessary to exhume large numbers of individual graves to include all unidentified immigrants. Ultimately the exhumation effort will significantly increase the probability of solving missing person cases.

In conclusion, the Reuniting Families database serves as an example of how modern informatics techniques can be brought together with applied anthropological efforts to address geographically distributed issues. It is the hope of the authors that as the database grows and additional state offices and families participate, there will be a significant reduction of the current 44% rate of unidentified deceased immigrants.

Acknowledgments

We would like to thank Dr. Bruce Anderson, Dr. Bruce Parks and Dr. Walt Birkby of the Pima County Medical Examiner's Office for their assistance and support of this project. We would also like to thank the many Mexican Consular officials that have invested many hours in helping us begin this work.

References

1. http://uscis.gov/graphics/shared/statistics/publications/msrsep04/SWBORD.HTM.
2. http://www.rtfcam.org/border/hilldrop080505.pdf.
3. Cornelius WA. Death at the border: efficacy and unintended consequences of US immigration control policy. Popul Dev Rev 2001;27(4):661–85.
4. http://www.rtfcam.org/border/hilldrop081605.pdf.
5. LoMonaco C. Many border deaths unlisted. The Tucson Citizen 2003 July 30; Sect. A:1 (col. 6).
6. PostgreSQL database homepage. 2005. http://www.postgresql.org.
7. The Apache Software Foundation. 2005. http://www.apache.org.
8. PHP: hypertext preprocessor. 2005. http://www.php.net.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
10. The Perl Directory. 2005. http://www.perl.org.
11. The Sun Developer Network. 2005. http://www.java.sun.com.
12. Stajich JE, Block D, Boulez D, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: perl modules for the life sciences. Genome Res 2002;12:1611–8.

Additional information and reprint requests:
Lori E. Baker, Ph.D.
Department of Anthropology, Forensic Science and Archaeology
Baylor University
One Bear Place #97173
Waco, TX 76798
E-mail: lori_baker@baylor.edu