Improving historical research by linking digital library information to a global genealogical database

Conference Paper · June 2009

DOI: 10.1145/1555400.1555441 · Source: DBLP

William B. Lund
Brigham Young University

Bryan S. Morse
Brigham Young University

Improving Historical Research by Linking Digital Library Information to a Global Genealogical Database

Douglas J. Kennard
Brigham Young University
Provo, Utah, U.S.A.
kennard@cs.byu.edu

William B. Lund
Brigham Young University
Provo, Utah, U.S.A.
bill_lund@byu.edu

Bryan S. Morse
Brigham Young University
Provo, Utah, U.S.A.
morse@cs.byu.edu

ABSTRACT

Journals, letters, and other writings are of great value to historians and those who research their own family history; however, it can be difficult to find writings by specific people, and even harder to find what others wrote about them. We present a prototype web-based system that enables users to discover information about historical people (including their own ancestors) by linking digital library content to unique PersonIDs from a genealogical database. Users can contribute content such as scanned journals or information about where items can be found. They can also transcribe content and tag it with PersonIDs to identify who it is about. Additional features provide tools for users to explore historical contexts and relationships. These include the ability to tag places and to create a historical social network by specifying non-family relationships or by using a mechanism we call *rosters* to imply participation in some group or event.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Standards, Systems issues, User issues*

General Terms

Design, Experimentation, Human Factors

1. INTRODUCTION

Glimpses into the lives, feelings, beliefs, and personalities of people who lived in the past can be experienced by reading their own journals, letters, and other writings. In addition, insights can often be gained from what others wrote about those people, even if the people wrote nothing about themselves. Similarly, reading first-hand accounts of those who participated in, witnessed, or were affected by historical events provides context and feeling that facts alone do not convey. These glimpses and insights enhance historical research in general and are particularly valuable to those who seek out their roots through family history research.

While the value of personal writings is easily recognized, it is difficult to actually know what (if anything) was written by a person, especially if the person was not considered noteworthy by historians. It is even more difficult to know what was written by others about the person.

Some writings have been donated to libraries or museums, but many have simply been passed down from one descendant to another without other descendants knowing, much less non-related historians. Even writings in libraries are unlikely to be found without prior knowledge that they exist and of the collection to which they were donated.

Online search engines are only minimally useful for finding journals, since most privately-held journals — and even many in libraries — are not reported online or indexed by search engines. Those that are may still be hard to find because of irrelevant search hits or ambiguity arising from common names such as “Maria Gonzalez” or “John Smith.”

We present a prototype web-based system that overcomes the difficult problem of discovering materials written by or about a particular person, enabling users to find journals, letters, or other materials relevant to the people that are of interest to them. While our system is primarily intended for scanned writings, the same principles are easily extended to other types of media and digital library (DL) artifacts.

Key to our system is allowing DL content to be tagged with identifiers that specify exactly which people the content pertains to, in effect eliminating ambiguity from the process of later searches. The *PersonID* identifiers used are from a very large genealogical database (Section 3.1). Since the database contains parent-child and marriage relationships, we can also search for writings about all ancestors of a user at once, instead of requiring a separate search for each ancestor. Tags using place identifiers can also be created to facilitate searches for writings pertaining to a specific location.

We provide a DL repository to which users may contribute artifacts such as scanned journals and letters. Users can collaboratively tag, transcribe, and annotate content. They can also add reference information for artifacts found elsewhere and add tags as if those artifacts were in the repository.

Users can create and explore historical social networks through *direct connections* (explicitly defined relationships) between people, and *implicit connections* (inferred by being listed on the same *rosters*). Rosters are lists of people who belonged to the same group (such as a military unit, church congregation, etc.) or who participated in the same event. People on the same roster may have been acquainted with each other, written about each other, or shared experiences that provide insight and context for users’ research.

2. RELATED WORK

Pera [2] uses word-correlation and folksonomies to improve catalog searches for library books, allowing intuitive search terms to be used instead of exact keywords or subject schemes defined by the Library of Congress. However, this would probably not help for personal journals since the journals would be unlikely to appear in the folksonomy used.

Crane and Jones [1] evaluate named entity analysis for automatic extraction of 10 entity classes in a newspaper collection. The *Personal Names* class accuracy is about 75%. The named entity analysis requires a full transcription of the materials, but many journals have not been transcribed, and handwriting recognition is not yet accurate enough for automatic transcription. Our system allows tagging of references to people in untranscribed handwritten journals.

Libraries and archives usually only have the resources to catalog those people who are major topics of an item or collection. Creating authority records to unambiguously identify people requires large amounts of time and effort, so it is rarely performed with enough detail to search for references to people who are not a main focus of the material. Users should be able to add information to aid searching without having to learn a strict, formal method of cataloging, an idea suggested in the literature and supported in a study by Sedgwick [3]. Our system allows users to do just that — PersonIDs from an existing genealogical database perform a role similar to authority records but with less effort since the records often already exist in in the database. Since many of the materials that are of interest to us are privately owned, the ability for users to contribute content and information that aids in searches is not only helpful but essential.

The Emmet[1], WeRelate[2], and WorldHistory[3] sites each provide some similar features as our system such as the ability to collaborate, set up historical social networks, and contribute source documents, photos, or artifacts. However, they are more biographical, depend on users to upload their genealogy, and are not well-suited for the specific task of searching for writings by or about people. We are not aware of other systems that use an existing genealogical database as an authority control framework for written DL artifacts.

3. SYSTEM DETAILS

3.1 FamilySearch Database and Internet API

The Church of Jesus Christ of Latter-day Saints (LDS Church) provides services and products that are well-known in the genealogical community such as the FamilySearch website[4], which provides free access to databases including Ancestral File, International Genealogical Index, and Pedigree Resource File. The databases include large amounts of user-contributed genealogical information submitted both by people who are and who are not members of the LDS Church, as well as data from extraction/indexing projects.

The Church is developing a new FamilySearch system[5] that makes it easier for users to collaborate and contribute additional information, as well as combine duplicate information and differing opinions of facts about an individual into a single record. People in the database are lineage-linked into a large family tree. Data from the previous databases is included, and each person in the new system has a PersonID (and additional "alternate" PersonIDs when records are combined).

The new FamilySearch system will eventually be available and free for everyone, and is already available to many LDS Church members. Over time, the interconnected family tree will grow, and likely converge to a relatively stable and accurate state as duplicate information is merged and errors are resolved. In addition to the web interface, an Internet API allows access to the database and other services (such as an extensive place authority) to approved projects that go through a certification process[6]. It is through this API that our system is able to access the FamilySearch database to update our lists of users' ancestors and to get lists of alternate PersonIDs by which a person can be referenced. To do so, users must authenticate by entering their FamilySearch username and password so that we can open a FamilySearch session and request data in their behalf through the API.

Since the FamilySearch database is constantly being updated, corrected, and added to by users of that system, our system permits users to synchronize their list of ancestors with FamilySearch at any time. The user's family tree is traversed for several generations and the PersonID (along with any alternate PersonIDs) of each ancestor is cached on our system for use during searching. Rosters (Section 3.4) can also be synchronized at any time.

3.2 User-Contributed Data

Users may contribute journals (or other materials) to our system or, alternatively, simply add reference information about where to find the materials. In either case, the user specifies who wrote a journal by looking up the author on the FamilySearch website and entering the PersonID of the author into our system. If the person does not exist in the database, the user can add the person on the FamilySearch website and then use the new PersonID.

Scanned images and transcriptions of journal pages may be added to our system at any time. Users may also collaborate with others and specify users who should have permissions to edit, tag, or transcribe the journal. Likewise, permission to view the journal can be granted to just some users, or the journal can be public for everyone to view.

3.3 Person and Place Tags

Tags specifying the people and places discussed in a journal may be added either by drawing a rectangle on the page image and entering the corresponding PersonID or place identifier (Figure 1), or by directly typing in the identifier and (optionally) the page number and coordinates. Tags may be specified even if the journal is not uploaded to the site, which allows searches to be performed based on the tags for journals held in other collections or that are not online.

3.4 Direct and Implicit Connections / Rosters

Users can specify direct connections between people to form a historical social network. For example, a user might create a direct connection between her ancestor and her ancestor's business partner by specifying both PersonIDs.

Users may not actually know who the acquaintances of their ancestors were, but know something that is useful in

[1] http://www.emmetlabs.com

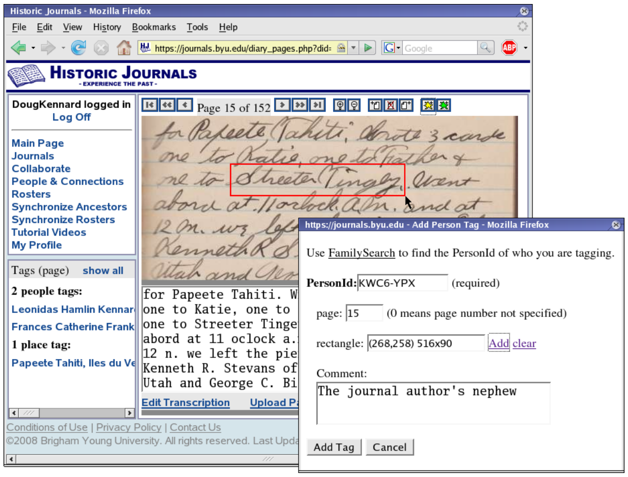
[2] http://www.werelate.org

[3] http://www.worldhistory.com

[4] http://www.familysearch.org

[5] https://new.familysearch.org

[6] https://devnet.familysearch.org



Historic Journals - Mozilla Firefox

File Edit View History Bookmarks Tools Help

https://journals.byu.edu/diary_pages.php?did=

HISTORIC JOURNALS

- EXPERIENCE THE PAST -

DougKennard logged in
Log Off

Main Page
Journals
Collaborate
People & Connections
Rosters
Synchronize Ancestors
Synchronize Rosters
Tutorial Videos
My Profile

Tags (page) show all

2 people tags:

Leonidas Hamlin Kennar
Frances Catherine Frank

1 place tag:

Papeete Tahiti, Iles du Ve

Page 15 of 152

for Papeete Tahiti, Wrote 3 cards
one to Katie, one to Father &
one to Streeter Tingey. Went
abord at 11 oclock a.m. and at
12 m. we left
Kenneth R S
Utah and Ge

for Papeete Tahiti. W
one to Katie, one to
one to Streeter Tinge
abord at 11 oclock a.
12 n. we left the pie
Kenneth R. Stevans of
Utah and George C. Bi

Edit Transcription Upload Pa

Conditions of Use | Privacy Policy | Contact Us
©2008 Brigham Young University. All rights reserved. Last Upda

https://journals.byu.edu - Add Person Tag - Mozilla Firefox

Use FamilySearch to find the PersonId of who you are tagging.

PersonId: KWC6-YPX (required)

page: 15 (0 means page number not specified)

rectangle: (268,258) 516x90 Add clear

Comment:
The journal author's nephew

Add Tag Cancel

Figure 1: Screenshot of the tagging interface

determining who they might have been. *Rosters* of people who participated in the same event or belonged to the same group can be created. People on a roster are considered to be *implicitly connected*, meaning they may have been acquainted, or at least shared common experiences that might interest those who research the lives of others on the roster.

For example, a user might know that his ancestor fought in Captain Stout's Army in the American Revolution. By placing his ancestor on a roster for that army, his ancestor is implicitly connected to any other people who are added to that roster. The user can easily find out if other members of that roster have journals that are recorded in the system.

The user who creates a roster decides whether other users may edit the roster, add people to it, or add sub-rosters.

3.5 Search

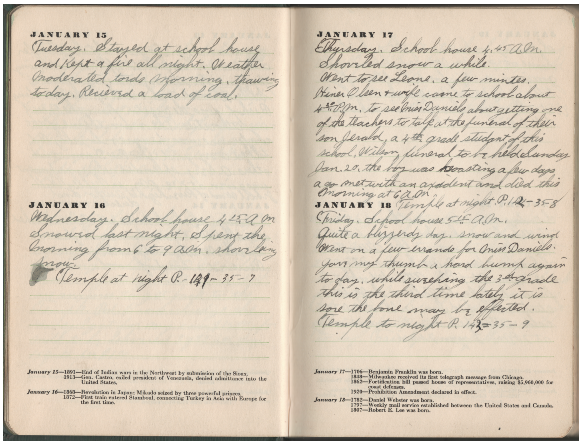
Perhaps the most useful part of our system is the search capability. Users can search for journals by a particular person (by PersonID) or for all journals of their own ancestors. More importantly, users can also search for tags that refer to any of their ancestors (or to the person they are searching for), as well as for any journals written by those people's direct connections or even implicit connections. This enables users to search for information written by others *about* their ancestors or other people of interest instead of just materials written by the people themselves.

3.6 Privacy and Ethical Considerations

Some genealogical information (e.g., birthdate, mother's maiden name) could potentially be abused for purposes like identity theft. The FamilySearch API does not permit users to access information of other living people, except for limited access to direct-line ancestors and descendants. To further prevent potential abuse, our own system does not cache living ancestors (even the user's direct line), and prohibits PersonIDs of living people to be used for tagging, marking journal authors, identifying roster members, or creating direct connections between people.

Due to the personal nature of journals, letters, and other materials our site is meant for, additional consideration of privacy is extremely important. Even if an author is dead, his or her journal may discuss people who are still living, or may divulge information about the author or other people that is inappropriate or that the people (or their descendants) would not want divulged even after their deaths.

Users must agree to policies that prohibit posting journals that give details about living people or that contain sensitive or otherwise inappropriate material, and are asked to err on the side of cautiousness if there is any question. Users are also encouraged to report any policy violations they find.



JANUARY 15

Tuesday. Stayed at school house
and kept a fire all night. Weather
moderated tords morning. thawing
today. Recieved a load of coal.

JANUARY 16

Wednesday. School house 4^{45} A.M.
Snowed last night. Spent the
morning from 6 to 9 a.m. shoveling
snow.
Temple at night P.-141-35-7

January 15—1891—End of Indian wars in the Northwest by submission of the Sioux.
1913—Gen. Castro, exiled president of Venezuela, denied admittance into the United States.

January 16—1868—Revolution in Japan; Mikado seized by three powerful princes.
1872—First train entered Stamboul, connecting Turkey in Asia with Europe for the first time.

JANUARY 17

Thursday. School house 4.45 A.M.
Shoveled snow a while.
Went to see Leone. a few mintes.
Hiner Olsen + wife came to school about
4^{30} P.M. to see Miss Daniels about getting one
of the teachers to talk at the funeral of their
son Jerald, a 4th grade student of this
school, Wilson funeral to be held Sunday
Jan. 20. the boy was coasting a few days
a go met with an axdent and died this
morning at 6 A.M.

JANUARY 18 Temple at night P.142-35-8

Friday. School house 5^{15} A.M.
Quite a blizzerdy day. snow and wind
Went on a few errands for Miss Daniels.
Gave my thumb a hard bump again
to day. while sweeping the 3rd grade
this is the third time lately it is
sore the bone may be effected.
Temple to night P. 143-35-9

January 17—1706—Benjamin Franklin was born.
1848—Milwaukee received its first telegraph message from Chicago.
1862—Fortification bill passed house of representatives, raising $5,960,000 for coast defenses.
1920—Prohibition Amendment declared in effect.

January 18—1782—Daniel Webster was born.
1797—Weekly mail service established between the United States and Canada.
1807—Robert E. Lee was born.

Figure 2: Pages of the case study journal

	Count	Pct	Page Avg
All References	90	100%	3.00
Unique Refs	61	68%	2.03
Family Refs	24	27%	0.80

Table 1: References to other people (all references, those that are not duplicate references to people already referenced, and references to relatives)

4. CASE STUDY

As a motivating example of the usefulness of our system, we examine part of a journal as a case study. The journal is that of a 65-year-old male living in Logan, Utah in 1935, who is an ancestor of the primary author of this paper. For each entry, the date is pre-printed, followed by several rule lines for the entry of that day. Half of a page is allocated for each entry, resulting in two entries per page (Figure 2).

First, we look at the number of descendants of the journal author. Using records updated in 2006 by a family genealogist, approximately 870 living descendants are counted, not including spouses. Due to births since 2006, we estimate that there are now well over 900 living descendants of the journal author. Most of the descendants are unaware of the existence of the journal, and even the author of this paper was unaware of it until recently, despite the fact that it had been in the possession of a first cousin for several years.

We now examine the contents of the first 30 journal pages. As shown in Table 1, the journal author makes reference to 90 people, averaging three references per page. 61 distinct people are mentioned, as some are referenced multiple times. 24 references (27%) are to known family members and the other 66 (73%) are to non-relatives.

Most references are counted individually, but in four cases we count multiple people as a single reference (for example, "The Handriks family" as one reference to a single person).

	People	Pct	Refs	Pct
Yes - Relatives (easy)	11	18%	24	27%
Yes - Others (easy)	12	20%	19	21%
Yes - Others (harder)	19	31%	28	31%
No / maybe	19	31%	19	21%

Table 2: Ability to tag case study references

Most references are by name, although a partial name or title is sometimes used, such as "Mrs. Wilcox," "Dr. Porter," or "Bp." (Bishop). In a few cases, not even part of a name is used. For example, one reference is to "the ward teachers," and another to "a man," who did some repair work.

In many cases, additional commentary disambiguates who people are, even when the reference would be ambiguous on its own. It would normally be difficult to know who "Miss Daniels" is, but context about her job at Wilson Elementary School permits us to find out enough about her through historical records to unambiguously tag the reference.

As shown in Table 2, PersonIDs for all 11 of the relatives (24 references) are easily located in FamilySearch for tagging purposes. We also easily locate PersonIDs for an additional 12 people (19 references) who are not relatives of the journal author. Most of these are found primarily using information found in the journal, sometimes combined with other information that is easily located online or that can be inferred due to basic knowledge of the family and community.

Information to tag 19 more people (27 references) is found through a moderate amount of additional investigation, such as comparing journal entries with events reported in the local newspaper, family and church records, and other history records for the area. One more reference to a previously identified person is also disambiguated (therefore, 28 total).

We do not tag 19 references. It may be possible to tag a few of them with additional effort and investigation, but some are unlikely to be tagged no matter how much effort is expended due to lack of contextual information.

PersonIDs already exist for all people that we tag (attributable to Logan's large LDS population), so we do not create any. For unmerged duplicate records, we could use any of the PersonIDs since records will eventually be merged by descendants (or we could merge them). For now, we use the PersonID from the record that is most accurate and has the most descendants listed, since it will be most useful until the records are merged. A perfectly clean database is not necessary for the system to be extremely useful.

We are pleasantly surprised by how many references we can tag in this journal. Ability to tag is affected by both the author's style and the availability of external information. It would certainly be more difficult to disambiguate people in a large city or where church, news, and other records are less detailed. However, even if only a handful of people could be identified, that would still be useful to those peoples' descendants or anyone else researching those people.

While writing styles, journal layout, entry lengths, and amount of detail vary greatly from one journal to another, this journal is not atypical of many others we have encountered in our research. A brief examination of five journals arbitrarily selected from the 376 digitized journals in the Mormon Missionary Diaries online collection[7] indicates that other diarists also make many references to people. Counting references made on 10 pages of each of the journals (50 pages total), we find an average of about seven references per page, with approximately 50% being unique references.

From this case study, we see that there are many people who could potentially have interest in the contents of this journal, even though its author would be considered a very ordinary person on most accounts. Not only are there over 900 living descendants of the journal author, but also the descendants of many other people he wrote about, who might also be interested in what he said about their ancestors. In only 30 pages (about one-sixth of the journal), he referenced 50 people (31 of whom we tag) who were not related to him. Although we do not know how many descendants the people he mentioned have, it is safe to say that the total of all their descendants would number at least in the thousands and probably in the tens or hundreds of thousands.

There is no way of knowing how many of those thousands of people actually would have interest in the journal, but the fact that there are so many people who potentially *could* have interest is certainly a compelling reason to suppose that our system will be useful to many people by allowing them to search for writings by and about their ancestors.

5. CONCLUSION

We have proposed a system for improving historical research by making writings by and about people easy to find. Key to our approach is using person identifiers from an existing global genealogical database as a replacement for an authority file. A prototype implementation allows users to contribute materials or reference materials found elsewhere, tag references to people and places, create historical social networks, and search for writings relating to specific people.

We have presented a case study which found that there are many people who could potentially be interested in even a single journal, considering the thousands of descendants of the author and the people he wrote about. A system such as ours should be useful to many people by allowing easier research into specific historical figures and allowing users to easily find writings by and about their own ancestors.

6. ACKNOWLEDGMENTS

This research is funded in part by the Gene Powell BYU Fund for Family History Research. We also thank Gordon Clarke and the Family History Department of The Church of Jesus Christ of Latter-day Saints for permitting us to use the FamilySearch API, William A. Barrett for some of the initial ideas and inspiration for this project, and the BYU Harold B. Lee Library for providing access to many journals.

7. REFERENCES

[1] Crane, G. and Jones, A. The challenge of Virginia Banks: An evaluation of named entity analysis in a 19th-century newspaper collection. In *JCDL'06*, 31–40, June 11-15, 2006.

[2] Pera, M. S. *Improving library searches using word-correlation factors and folksonomies.* Master's thesis, Brigham Young University, 2009.

[3] Sedgwick, J. M. *Let me tell you about my grandpa: A content analysis of user annotations to online archival collections.* Master's thesis, University of North Carolina at Chapel Hill, 2008.

[7] http://www.lib.byu.edu/dlib/mmd/index.html