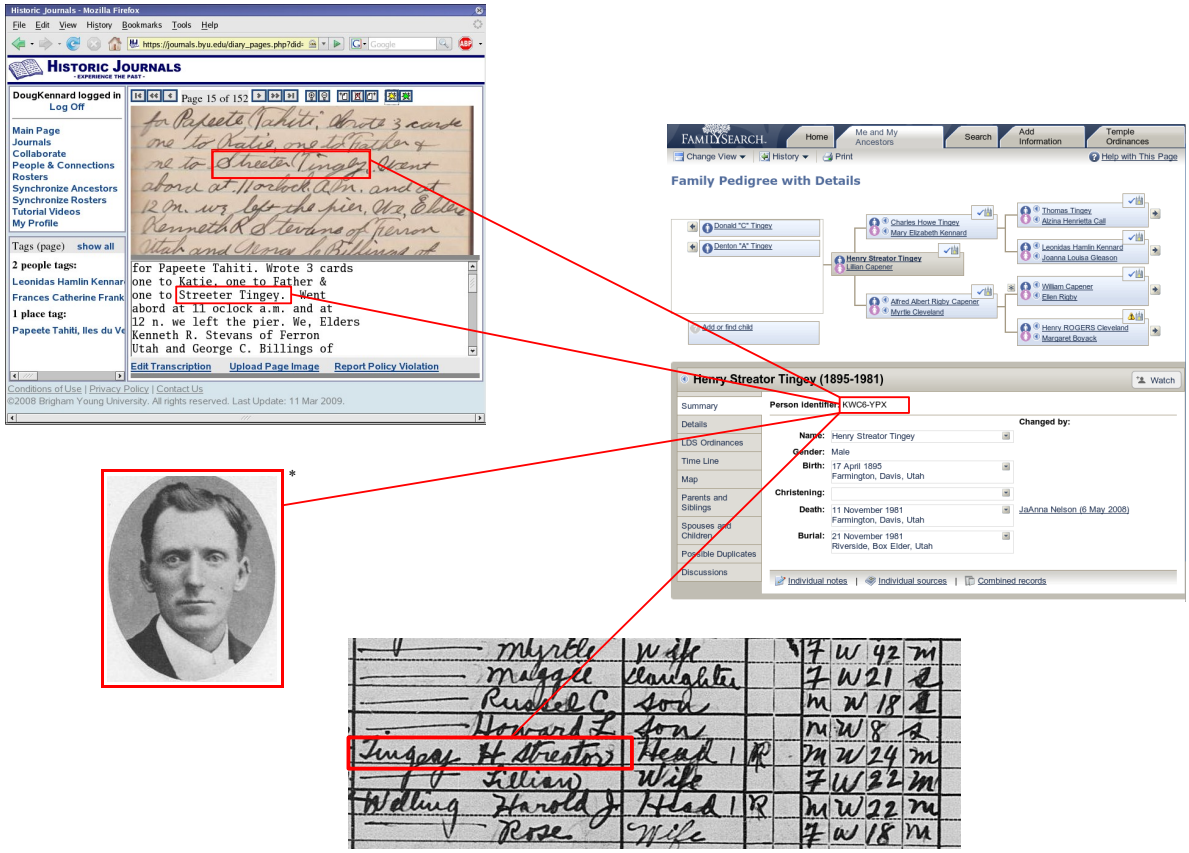
Record Linkage and Tagging for the BYU Historic Journals Project (journals.byu.edu)

Douglas J. Kennard
kennard@cs.byu.edu

William A. Barrett
barrett@cs.byu.edu

Abstract

We describe techniques used to link resources to new FamilySearch PersonIDs in the BYU Historic Journals Project. We describe both manual (crowd-sourced) tagging, and automatic/semi-automatic record linkage. We also describe possible future research directions.



1. Introduction

The BYU Historic Journals Project (journals.byu.edu) is proof-of-concept system that shows that it is very useful to tag resources such as diaries, letters, and photographs directly with the new FamilySearch PersonIDs of the subjects of those resources. This technique, which we presented in [1], allows users to easily find all items in the system pertaining to any of their ancestors automatically instead of having to perform numerous keyword or name searches, which are often inaccurate and

*The example photo is not actually Streator Tingey.

ambiguous. A user's pedigree is downloaded from new FamilySearch and then the PersonIDs of their ancestors are compared with the items in the system to see if any diaries were written by any of their ancestors, or even if ancestors have been tagged in diaries, photos, etc. of other people who may not even be related. The system also allows historical social networks to be created.

We previously described the system itself in more detail in JCDL 2009 [1]. In this paper we describe some of the methods we use (and some that we anticipate using in the future) for actually performing record-linkage. While we use these methods specifically for our system, the same techniques can be applied by others.

2. Record Linkage Methods Used

One of the great strengths of a system such as the Historic Journals Project is the ability for people to leverage each others' work to find information about ancestors. However, until a critical mass of existing data is put into the system, most people probably will not find it very useful. In order to boot-strap the system, we use several methods to tag and link to some existing data sources.

2.1 Manual Tagging / Crowd-Sourcing

The BYU Harold B. Lee Library has online collections of diaries written by Mormon missionaries and pioneers (http://www.lib.byu.edu/dlib/mmd/ and http://overlandtrails.lib.byu.edu/). We determine the PersonIDs of the authors of those diaries manually by using the biographical information in the collections and the search functionality of the new FamilySearch web application. The image viewer of our system allows all users to add tags of the people discussed in the diaries.

Davis Bitton's Guide to Mormon Diaries and Autobiographies is a list of 2,894 known diaries with short descriptions of the content matter, some biographical information about the author (if known), and the location of the diary. We manually use the information contained in the guide, sometimes disambiguating between people by cross-referencing with data found in the online pioneer database (http://classic.lds.org/churchhistory/library/pioneercompanysearch/1,15773,3966-1,00.html). The searches are performed manually through the new FamilySearch web application in order to tag the guide.

2.2 Automatic / Semi-automatic Record linkage

Pioneers and Prominent Men of Utah written in 1913 by Frank Esshom is a reference book with photographs, genealogies, and biographical information about several thousand early settlers of Utah. A scanned version (including OCR text) made available by the BYU Harold B. Lee Library is used to extract photographs and associate them with their corresponding PersonIDs.

The format of the photo pages is fairly consistent (see Figure 1). Each photo is automatically extracted, and obvious cropping errors are corrected manually. The OCR text is also manually corrected for obvious errors, since the OCR accuracy is poor. The OCR text is then parsed for the person's name, and other information that may be present such as birthdate, birth place, and parents' names. The FamilySearch API is used to automatically perform searches for people who match the criteria, and then up to three PersonIDs that match with high confidence are associated with the photograph. Minor errors are often tolerated by the automatic search, so the OCR correction does not need to be perfect.

Figure 1: Photograph pages from *Pioneers and Prominent Men of Utah* (expired copyright).

The Mormon Overland Travel pioneer database provides rosters of pioneer companies along with known birth and death dates. It is publicly available online at the URL (http://classic.lds.org/churchhistory/library/pioneercompanysearch/1,15773,3966-1,00.html).

We show proof-of-concept of how the people on the pioneer rosters can be automatically associated with their corresponding PersonIDs. We crawl a single roster and scrape the name, birth date, gender, and death date of each person on the roster. We then perform an automatic search using the FamilySearch API – similar to the search performed with the pioneer photographs – to determine the PersonIDs of the people on the roster. This same process could be used to automatically associate the entire database with PersonIDs. We do not actually add the data to the Historic Journals website since we have not received permission to do so.

3. Future Directions

We anticipate investigating automatically linking to PersonIDs for additional types of documents such as birth, death, and census records that are already transcribed. Those documents have information similar to the information we used for linking to PersonIDs for photographs and pioneer company roster members (name, birth/death dates, and sometimes parents' names).

We may also perform semi-automatic tagging of people mentioned in diary transcriptions. By automatically finding the names of people in a diary, comparing the names to the diary author's family tree and other resources (pioneer rosters the person belonged to, city directories, census information of the time and place, etc.) a tagging tool should be able to present a user with very good suggestions of who the people mentioned in the diary are, along with the corresponding PersonIDs for each suggestion.

References

[1] Douglas J. Kennard, William B. Lund, and Bryan S. Morse. "Improving Historical Research by Linking Digital Library Information to a Global Genealogical Database," Joint Conference on Digital Libraries (JCDL), 2009, Austin, Texas.