Word-Spotting for Automatic Tag Suggestion In the BYU Historic Journals Project

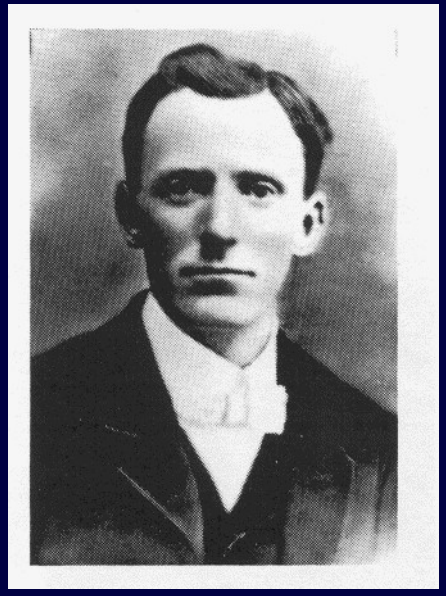
Douglas J. Kennard and Dr. Bryan S. Morse
(BYU Computer Science Department)

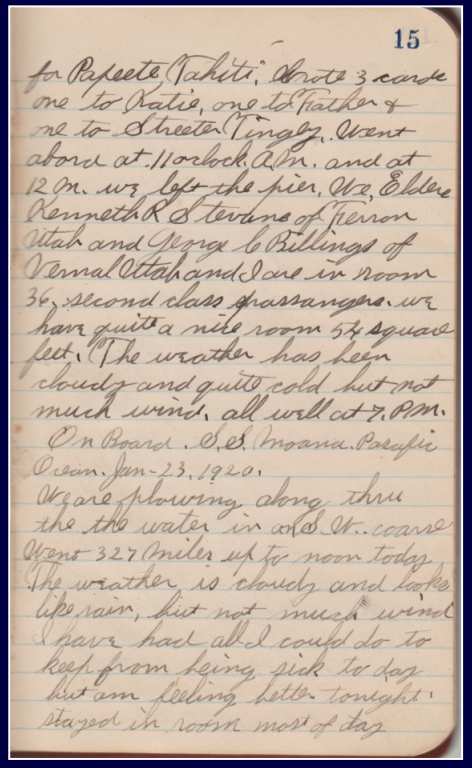
Journals, Letters, other writings

- **Open the door for interest in Family History**

- **Bring ancestors to life**

- **Help us understand / appreciate them**





15

for Papeete Tahiti. Wrote 3 cards
one to Katie, one to Father &
one to Streeter Tingey. Went
aboard at 11 o'clock A.M. and at
12 M. we left the pier. Wm. Elder
Kenneth R Stevens of Heron
Utah and Jesse C Billings of
Vernal Utah and I are in room
36, second class passengers. We
have quite a nice room 5x7 square
feet. The weather has been
cloudy and quite cold but not
much wind. all well at 7 P.M.

On Board S.S. Moana. Pacific
Ocean. Jan-23, 1920.

We are plowing along thru
the the water in a S.W. course
Went 327 miles up to noon today.
The weather is cloudy and looks
like rain, but not much wind
I have had all I could do to
keep from being sick to day
but am feeling better tonight.
stayed in room most of day

Problem: Journal Access

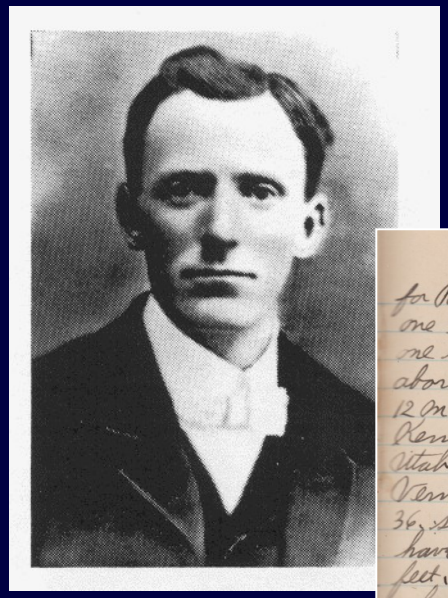
Did he have a journal?

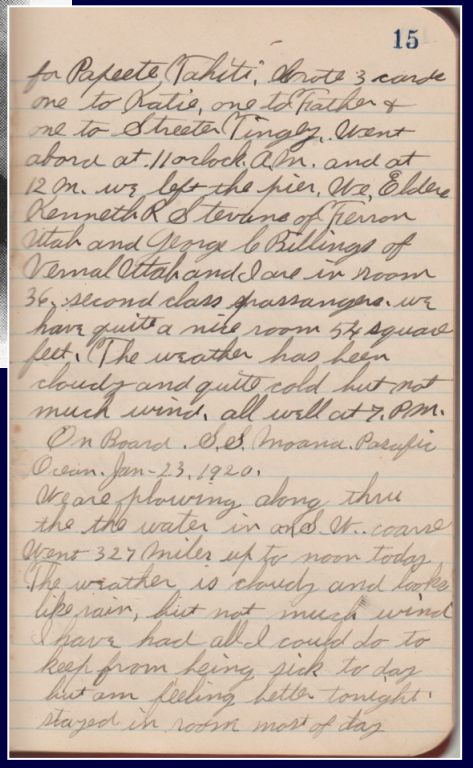
Does it still exist?

Who has it?

(1 of 900 living descendants)

How can I read it?





15

for Papeete Tahiti. Wrote 3 cards one to Katie, one to Father & one to Streeter Tingey. Went aboard at 11 o'clock A.M. and at 12 M. we left the pier. Pres. Elder Kenneth R. Stevens of Heron Utah and Pepsa C Billings of Vernal Utah and I are in room 36, second class passengers. we have quite a nice room 5x7 square feet. The weather has been cloudy and quite cold but not much wind. all well at 7. P.M.

On Board S.S. Moana. Pacific Ocean. Jan-23. 1920.

We are plowing along thru the the water in old St. coarse Went 327 miles up to noon today. The weather is cloudy and looks like rain, but not much wind I have had all I could do to keep from being sick to day but am feeling better tonight. stayed in room most of day

Has anyone else written ABOUT him?

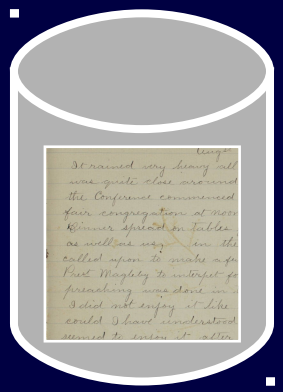
Did he write about others?

(their descendants would want to know)

Do my other ancestors have journals?

BYU Historic Journals Project

Search for writings by or <u>about</u> ancestors



API

PersonIDs (FamilySearch)

Submit Journals
- Scanned Images
- Transcriptions
- Reference Information

Collaboration
- Transcribe (if desired)
- Tag names (link to PersonIDs)
- Tag places (place authority)

BYU Historic Journals Project

Policies to protect privacy, avoid embarrassing:

- **living descendants**
- **ancestors**

“And that same sociality which exists among us here will exist among us there...”

(D&C 130:2)

System Details

Joint Conference on Digital Libraries (JCDL 2009)

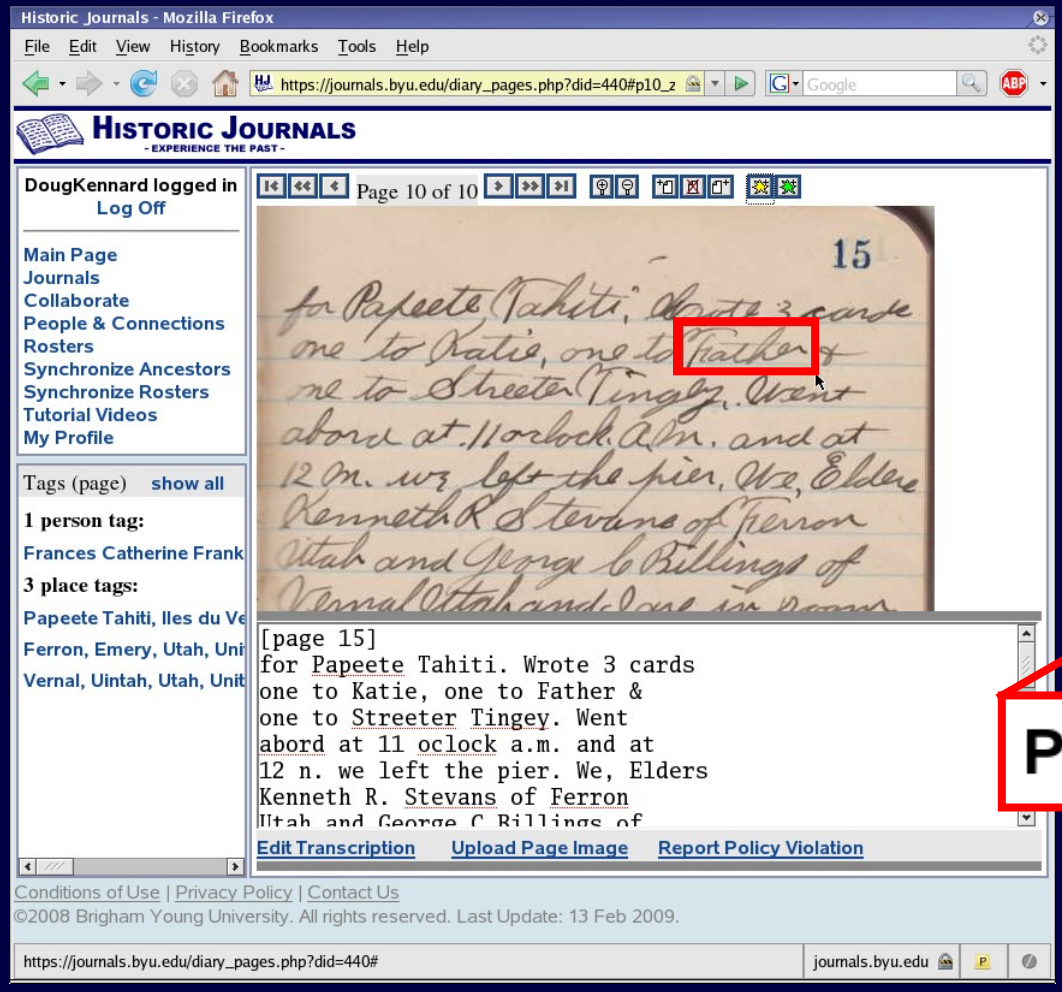
Douglas J. Kennard, William B. Lund, and Bryan S. Morse. “Improving Historical Research by Linking Digital Library Information to a Global Genealogical Database.” (to appear) JCDL 2009, Jun 15-19, 2009, Austin, TX.

Demo / Q&A: today in demo session

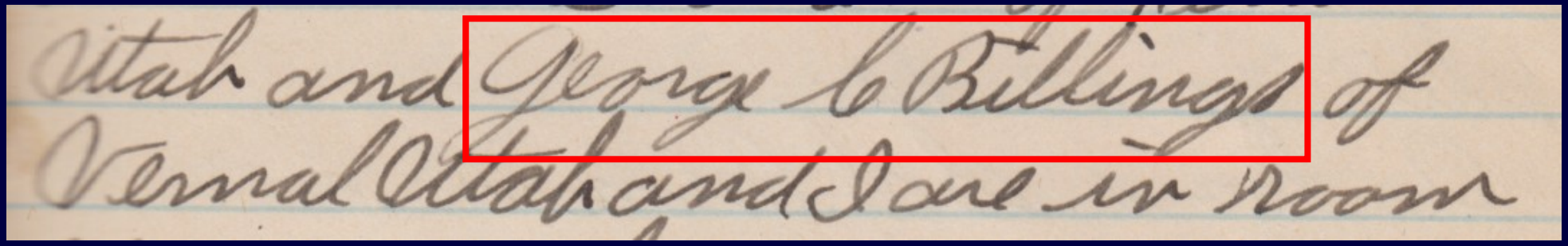
This Presentation

Word-spotting tools to aid with tagging

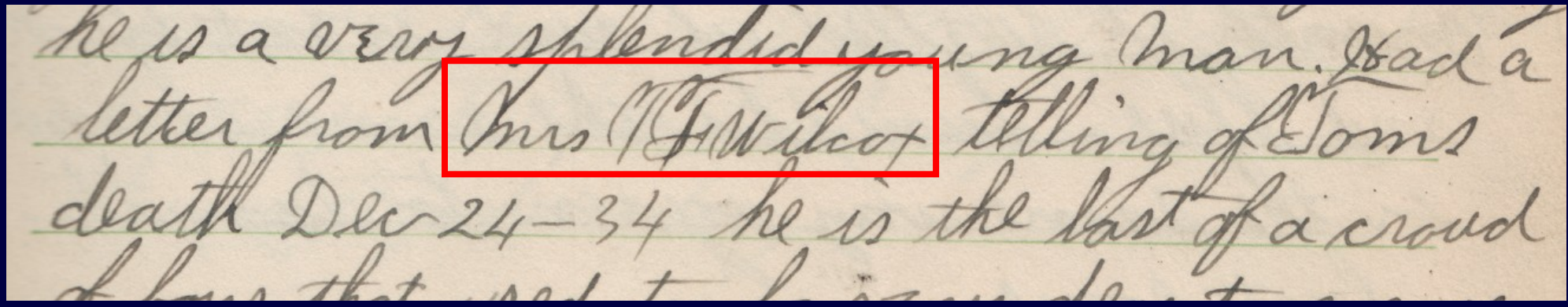
The Tagging Process



The Tagging Process



“George C Billings of Vernal Utah” - easy



“Mrs. TF Wilcox” - additional context

Observation

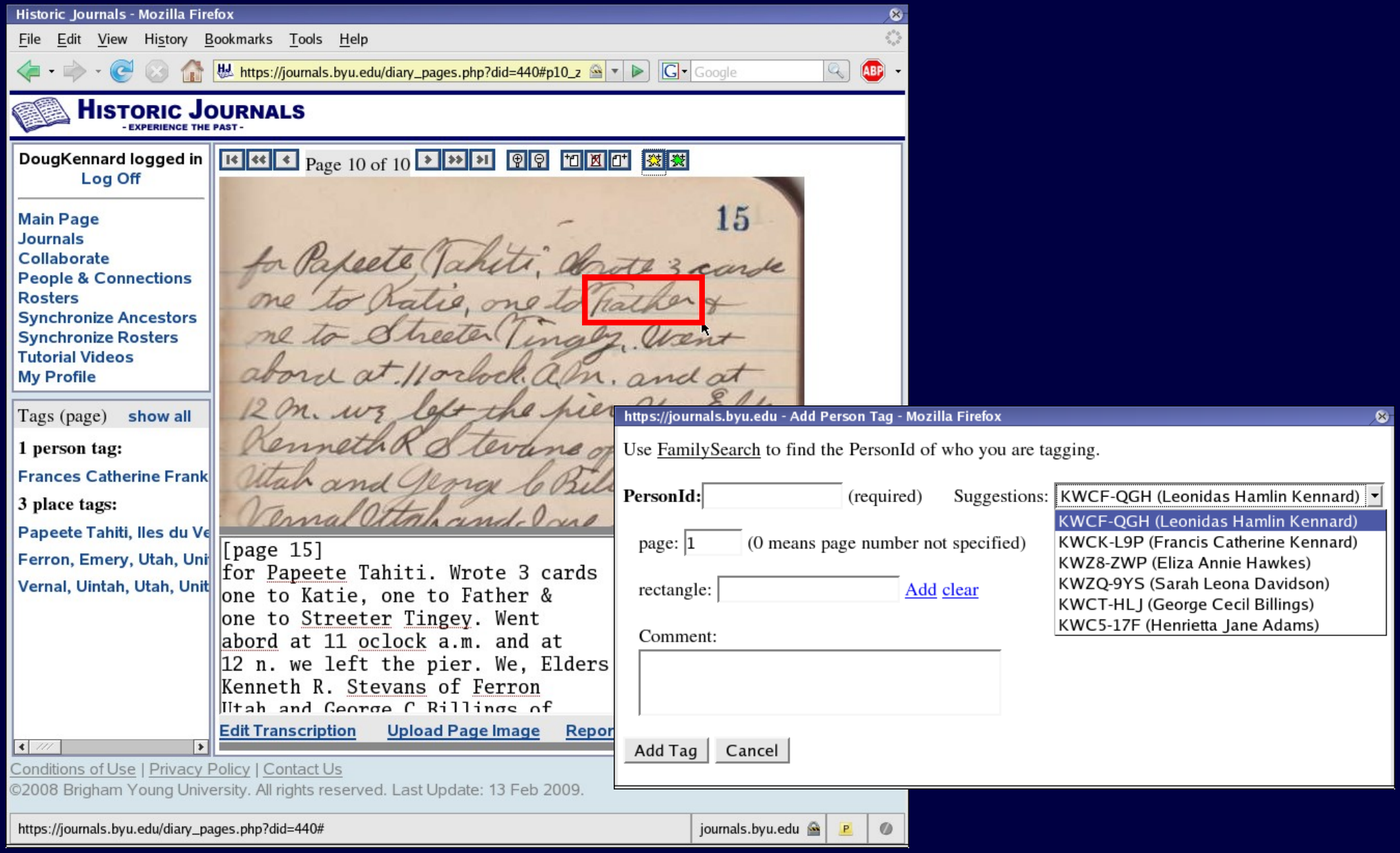
We often tag the same people many times

- look them up the PersonID again

- look back through our list of tags

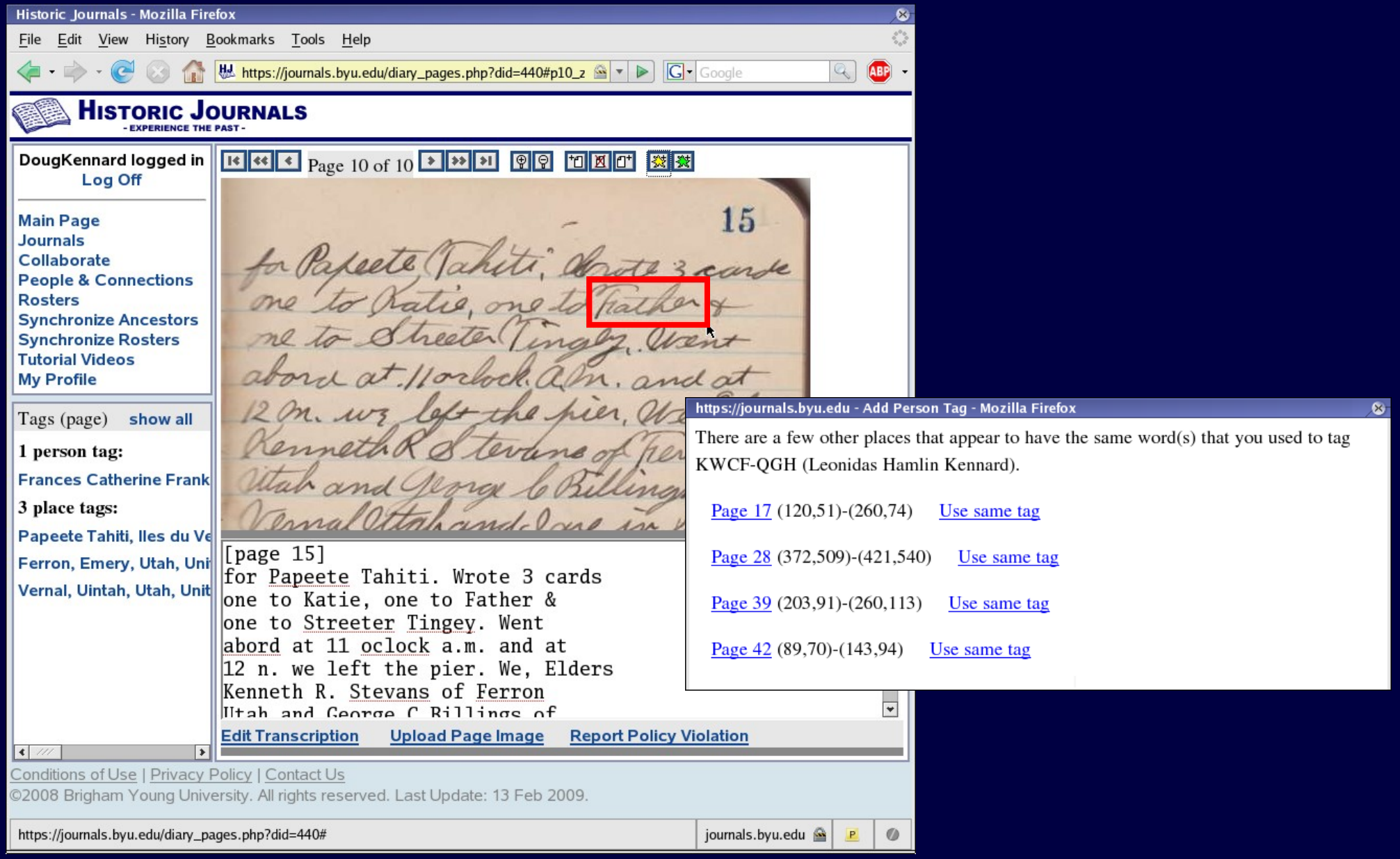
(neither is difficult, but both are inconvenient)

Proposed Tools



1- Suggest previous tags (order by similarity)

Proposed Tools



2- For a tagged word, spot other occurrences

Proposed Tools

Word-spotting:

- **We don’t need high accuracy (unlike transcription)**
- **Just find words that look similar (simpler problem)**
- **Mistakes are tolerable (user selects)**

Current Status

- Just getting started (for this application)

- Leverage code from our previous HR efforts:

Kennard and Barrett. “Progress with Searchable Indexes for Handwritten Documents,” (FHT 2007).

Methods

Preliminary Processing (offline):

1- Preprocessing (clean image, find ink)

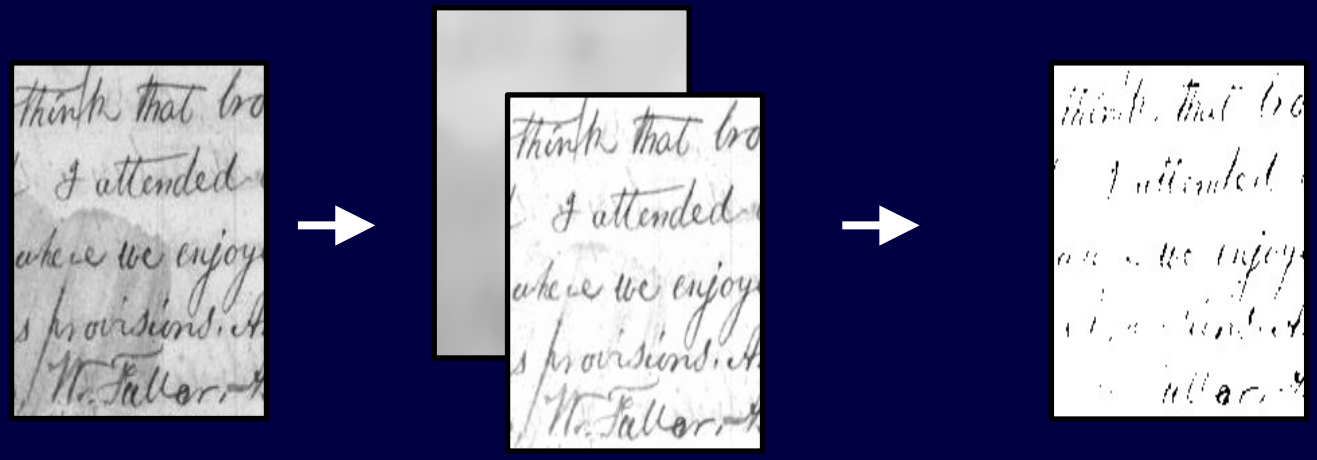
2- Segmentation (lines of text, words)

3- Feature Extraction

4- Save features for later use

1 - Preprocessing

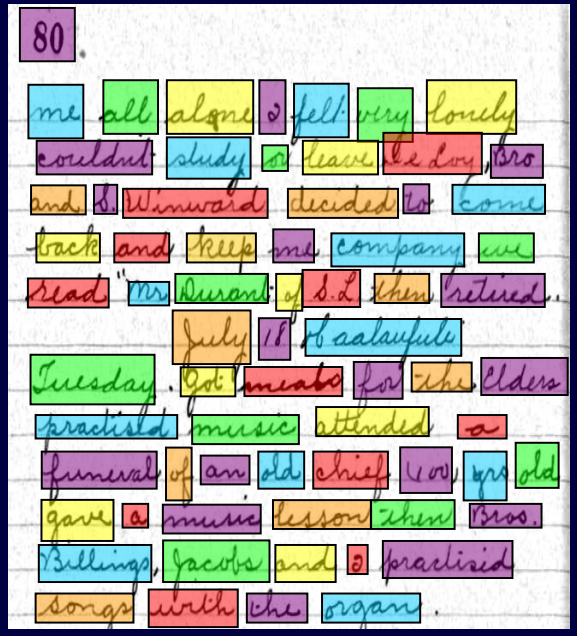
- **Clean image (filter noise)**
- **Remove borders, background**
- **Binarize image (find ink)**



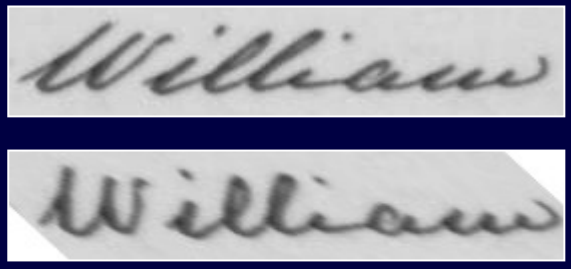
2 - Segmentation

- separate lines of text (profiles)

- word separation (gap metrics)



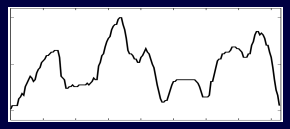
Derived from a page from Jennie Leavitt Smith's Diary.
Original imageFrom "Mormon Missionary Diaries."
Harold B. Lee Library, Brigham Young University, online collection,
available at http://www.lib.byu.edu/dlib/mmd



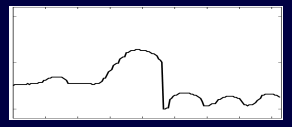
slant removal (shear)

3 - Feature Extraction

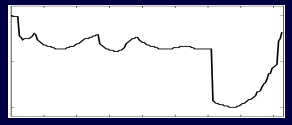
(Rath, Manmatha, Lavrenko, SIGIR 2004.)



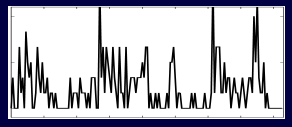
Projection Profile



Upper Profile



Lower Profile



Transition Counts

- Treat each as a 1-D signal

- Compute Fourier Transform

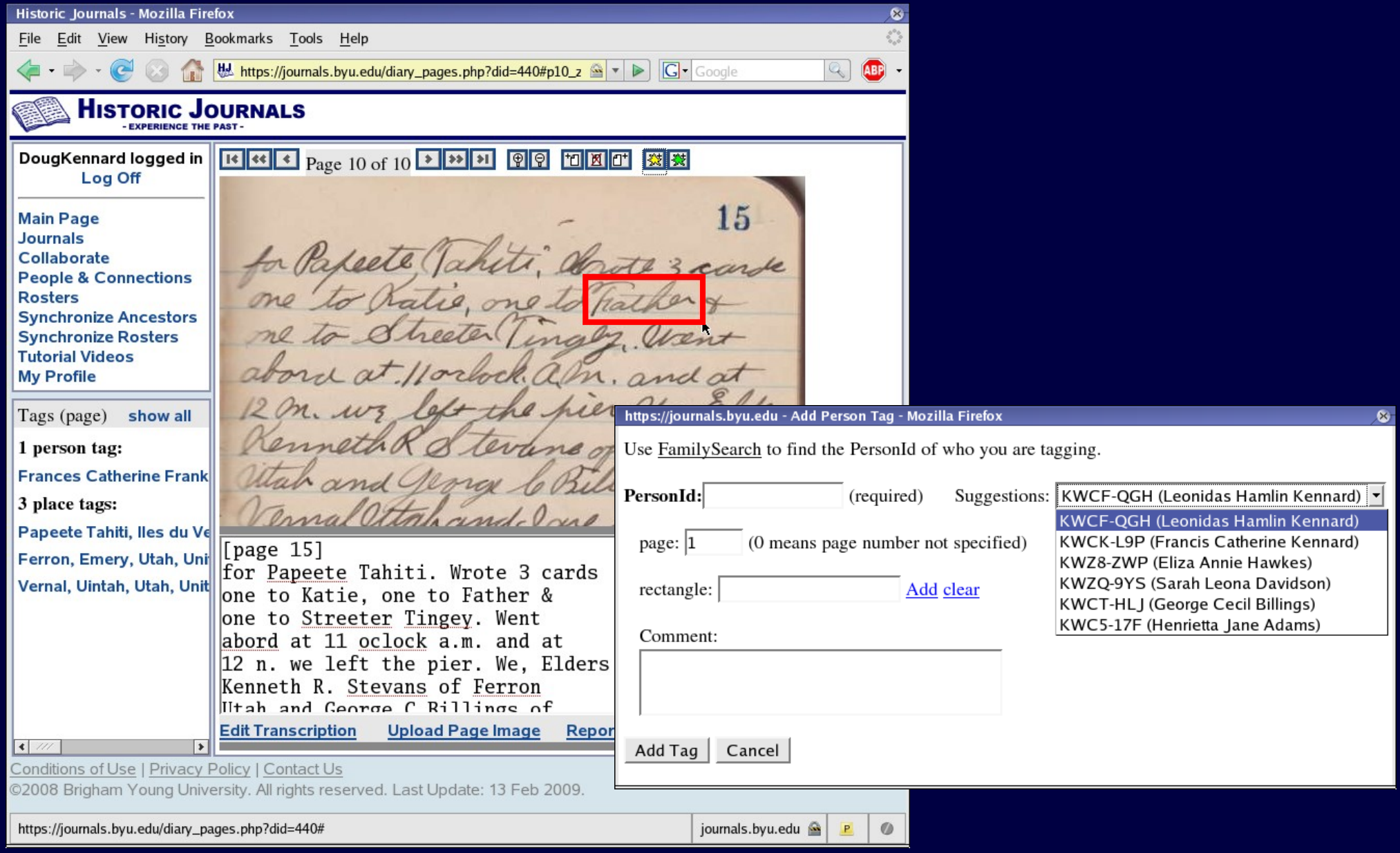
- Store Low-order Fourier Coefficients

4 - Save Feature Vectors

- Save feature vector of each word

For two words, their “difference” is calculated as Euclidean Distance between their feature vectors

Real-time Tag Suggestion



Conclusion

- **We proposed tools to help users tag journals:**
  - **Real-time Tag Suggestion**
  - **Search for other occurrences of Tag Words**
- **The tools are based on word-spotting**
- **We are currently in early stages of the research**
- **We hope tools will increase convenience for users**

Thank You