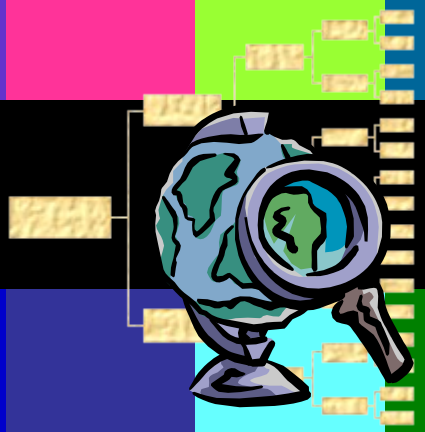


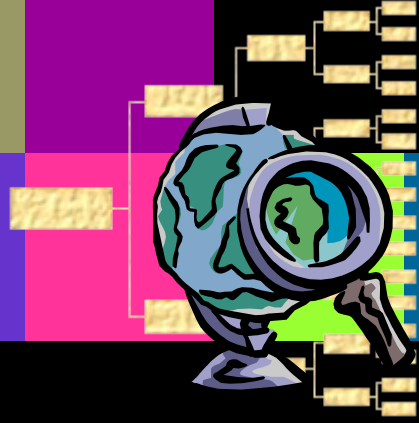
Genealogical Place Name Normalization

Bob Leaman
(bob.leaman@asu.edu)

What is meant by “Normalization”?

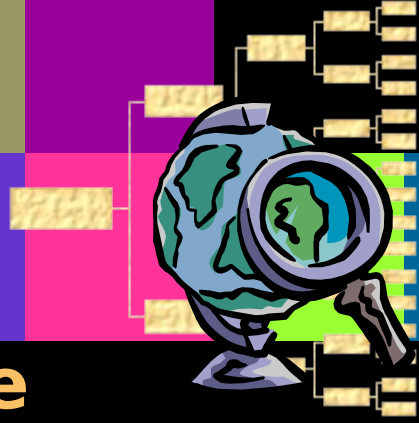
- **Enforcing a standardized representation**
- **Increases accuracy**
  - Data shared over e-mail can be very hard to correct
  - Easier record linkage
    - Automated merging
    - Automated research

What format to use?



- **Fixed three-level**
  - *Mesa, Maricopa, Arizona*
- **Variable-level**
  - *Mesa, Maricopa, Arizona, United States*
- **Note absence of descriptors**
  - “Of”, “Near”, etc.

The Problem



What kinds of deviations from the standard are common?

- **Biographical notes**
  - *Johnsville, Arkansas. He had 6 children*
- **Addresses and e-mails**
- **Hospital, church and cemetery names**
  - *Bluff Cemetery, Elgin, Ill.* → *Elgin, Ill.*
- **Leaving out one or more of the levels**
  - *Vancouver, Washington* → Vancouver, Clark, Washington, United States

The Problem

- **Excluding the comma between two of the place names**
  - *San Leandro CA* → *San Leandro, CA*
- **Using an abbreviated, truncated, or alternate form of a place name**
  - *UT* → Utah
  - *Tenn* → Tennessee
  - *Holland* → Denmark
- **Misspelling place names**
  - *Ypfilanti, Washtinaud, Michigan* → Ypsilanti, Washtenaw, Michigan
- **Algorithmic contractions such as removing all vowels after the first letter**
  - *Oxfrd* → Oxford

Strategy

- Preprocessing – remove everything that is not part of the place name
- Match against a name variations database (thesaurus)
- Match against standardized names database (gazetteer)

Preprocessing Place Names

- **Use regular expressions to detect patterns**
  - 38th year, Benedict, Kansas. Buried High Prairie Cem, Wilson, Kansas
    becomes
  - 38th year, Benedict, Kansas.
    becomes
  - Benedict, Kansas
- **List of “note words” (e.g. occupations, causes of death, etc.)**

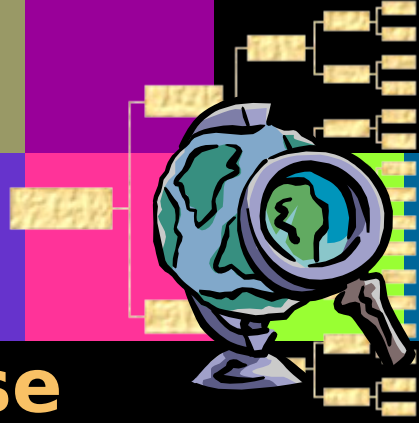
Preprocessing Place Names

- Tested on 2450 randomly selected “PLAC” fields from 10 different GEDCOM files
- Each was preprocessed by hand: 58.4% required modification
- Preprocessing via the system matched preprocessing by hand 97.6% of the time

Handling Name Variations

- **At this point all non-place name information has been removed**
- **Each place name is looked up in a database of alternate names (thesaurus)**
  - *Livonia, MI* → {Livonia, MI & Livonia, Michigan}
- **The original is included in case the wrong alternate was recorded originally**

Place Name Matching

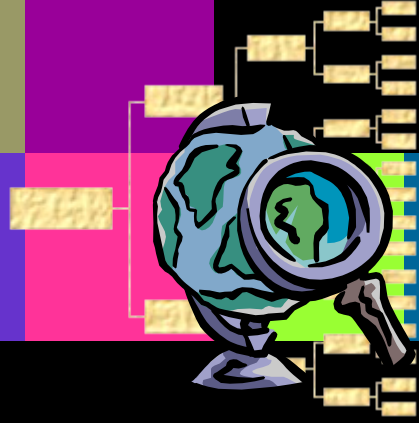


- **Created a place name database**
  - Mostly GNIS data
  - Includes all of the United States and some of England and Canada
  - Nearly 160,000 places
- **Database format**
  - A single table was used to hold all place records
  - Utilized unique identifiers to point to the “parent” record

Place Name Matching

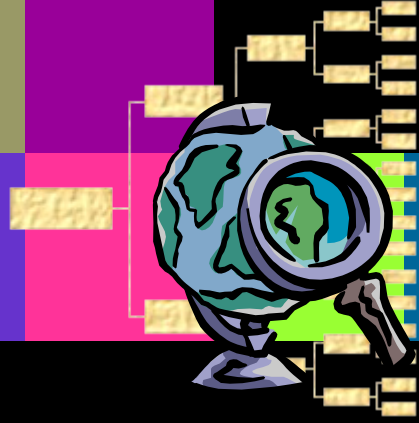
- **Need to find the place name in the database that maximizes the "similarity" with respect to the input place name**
  - 0 = no match
  - 1 = perfect match
- **Calculated using the average "similarity" of the individual pieces of the place name**

Place Name Matching



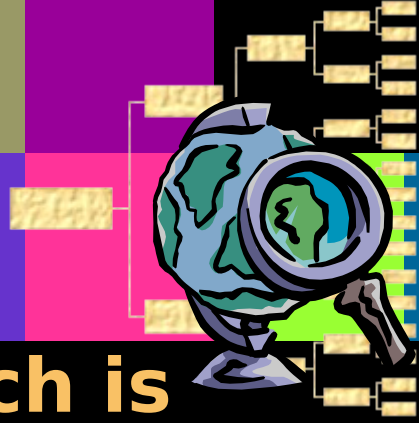
- **Used the elements of the edit distance metric**
  - Substitution, insertion, deletion
  - Added transposition, length of the longest common substring & a measure of truncation
- **Sorted through the several data points per potential match with a decision tree**
  - Trained using the metric scores from a test set of place name pieces matched by hand
    - *S Lk, Salt Lake, TRUE*
  - Used the proportion of test cases that were matches in any leaf of the tree as the "similarity" score

Place Name Matching



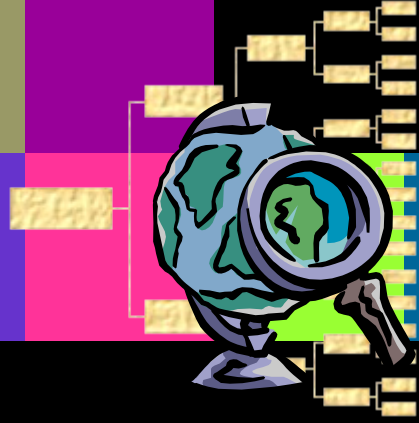
- Tested on 330 randomly selected “PLAC” fields from 10 different GEDCOM files
- Each was preprocessed and matched by hand: 99.1% required modification after preprocessing
- The first-ranked match was the same as the match found by hand 97.9% of the time
- The average rank of the match generated by hand was 1.21

Future Directions



- **Recognize when the best match is not satisfactory**
- **Acquisition of a suitable thesaurus and gazetteer**
  - Alexandria Digital Library Project
- **Historical place information**
- **Increased productization**
  - Indexing scheme
- **Internationalization**

Questions?



- **Reference:**

  K. Kukich. *Techniques for Automatically Correcting Words in Text.* Computing Surveys, 24(4):377-440, Dec. 1992.