# ELIJAH: Extracting Genealogy from the Web

By
**David Barney**
and
**Rachel Lee**

***WhizBang! Labs***

# Introduction

"A new era of family history work has arrived. As President Gordon B. Hinckley recently noted, 'The Lord has inspired skilled men and women in developing new technologies which we can use to our great advantage in moving forward this sacred work.' "

Elder Russell M. Nelson, "A New Harvest Time," Ensign, May 1998, 43

WhizBang! LABS

# Introduction: The General Problem

- There is a large amount of genealogical information already published on the web.
- How do you put it into a usable format?
- A search engine would be nice.

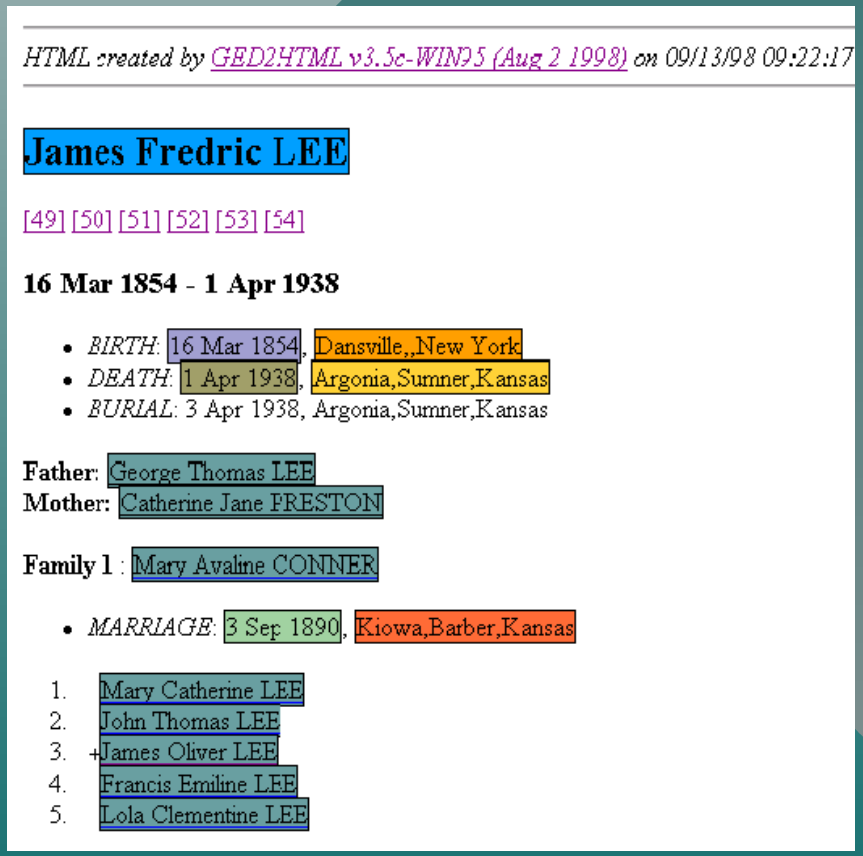
WhizBang! LABS

# Introduction: The Specific Problem

- Key word search is not good enough.
  - Is 1897 a death date, birth date, etc. ?
- 2 main problems with extracting information
  - Finding the fields (names, birthdates…)
  - Associating the fields into records

WhizBang! LABS

# Example: a Genealogy Page



*HTML created by GED2HTML v3.5c-WIN95 (Aug 2 1998) on 09/13/98 09:22:17*

**James Fredric LEE**

[49] [50] [51] [52] [53] [54]

**16 Mar 1854 - 1 Apr 1938**

- *BIRTH*: 16 Mar 1854, Dansville,,New York
- *DEATH*: 1 Apr 1938, Argonia,Sumner,Kansas
- *BURIAL*: 3 Apr 1938, Argonia,Sumner,Kansas

**Father**: George Thomas LEE
**Mother**: Catherine Jane PRESTON

**Family 1** : Mary Avaline CONNER

- *MARRIAGE*: 3 Sep 1890, Kiowa,Barber,Kansas

1. Mary Catherine LEE
2. John Thomas LEE
3. +James Oliver LEE
4. Francis Emiline LEE
5. Lola Clementine LEE

HTML page

| | |
|---|---|
| Individual Name | James Fredric LEE |
| Birth Date | 16 Mar 1854 |
| Birth Place | Dansville,,New York |
| Marriage Date | 3 Sep 1890 |
| Marriage Place | Kiowa,Barber,Kansas |
| Death Date | 1 Apr 1938 |
| Death Place | Argonia,Sumner,Kansas |
| Name | George Thomas LEE<br>Catherine Jane PRESTON |

Relational/XML Database

WhizBang! LABS

# Related Work: Wrappers

- Make a site-specific set of rules
- Pro: highly accurate
- Cons: not scalable, fragile

WhizBang! LABS

# Related Work: Global Models

- General approach
  - example: ***FlipDog.com***
- Pros: applies to any website, scalable
- Cons: time consuming to train/tune, possible to have low accuracy on specific sites

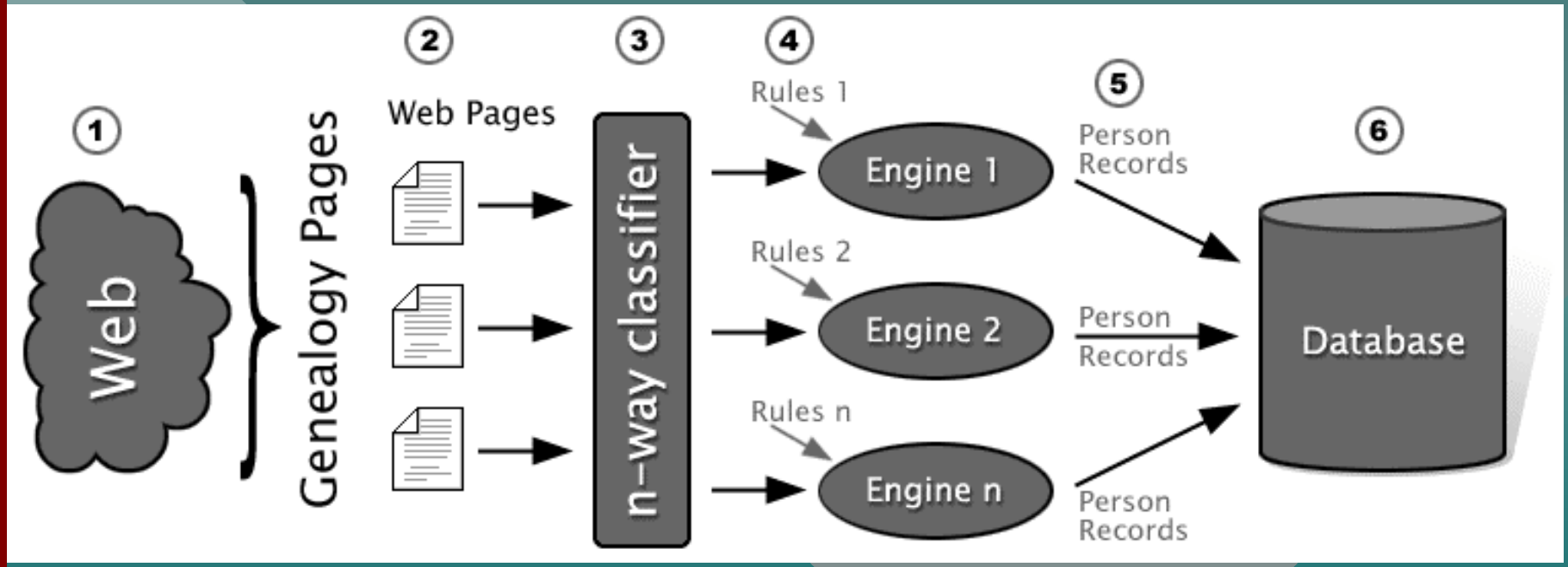
WhizBang! LABS

# Our approach: ELIJAH

- Key: 1000s of pages are produced by about 100 different software programs.
- Combines the two previous methods
- **E**xtracting **L**ineage **I**nformation with **J**ava using **A**utomated **H**euristics

WhizBang! LABS

# ELIJAH Architecture



# Example: ELIJAH in action


Ged2HTML rules

*HTML created by GED2HTML v3.5c-WIN95 (Aug 2 1998) on 09/13/98 09:22:17*

**James Fredric LEE**

[49] [50] [51] [52] [53] [54]

**16 Mar 1854 - 1 Apr 1938**

- *BIRTH*: 16 Mar 1854, Dansville,,New York
- *DEATH*: 1 Apr 1938, Argonia,Sumner,Kansas
- *BURIAL*: 3 Apr 1938, Argonia,Sumner,Kansas

**Father:** George Thomas LEE
**Mother:** Catherine Jane PRESTON

**Family 1** : Mary Avaline CONNER

- *MARRIAGE*: 3 Sep 1890, Kiowa,Barber,Kansas

1. Mary Catherine LEE
2. John Thomas LEE
3. +James Oliver LEE
4. Francis Emiline LEE
5. Lola Clementine LEE

*HTML created by GED2HTML v3.5c-WIN95 (Aug 2 1998) on 09/13/98 09:22:17*

**James Fredric LEE**

[49] [50] [51] [52] [53] [54]

**16 Mar 1854 - 1 Apr 1938**

- *BIRTH*: 16 Mar 1854, Dansville,,New York
- *DEATH*: 1 Apr 1938, Argonia,Sumner,Kansas
- *BURIAL*: 3 Apr 1938, Argonia,Sumner,Kansas

**Father**: George Thomas LEE
**Mother:** Catherine Jane PRESTON

**Family 1** : Mary Avaline CONNER

- *MARRIAGE*: 3 Sep 1890, Kiowa,Barber,Kansas

1. Mary Catherine LEE
2. John Thomas LEE
3. +James Oliver LEE
4. Francis Emiline LEE
5. Lola Clementine LEE

| | |
|---|---|
| Individual Name | James Fredric LEE |
| Birth Date | 16 Mar 1854 |
| Birth Place | Dansville,,New York |
| Marriage Date | 3 Sep 1890 |
| Marriage Place | Kiowa,Barber,Kansas |
| Death Date | 1 Apr 1938 |
| Death Place | Argonia,Sumner,Kansas |
| Name | George Thomas LEE<br>Catherine Jane PRESTON |

WhizBang! LABS

# Experiment

- Rules for 15 most common formats (out of 100)
- Executed ELIJAH on 51 random websites with family tree information
- Failed if
  - couldn’t identify what format it was
  - didn’t extract information
  - extracted information had errors

WhizBang! LABS

# Results

- With the 15 rule sets, we extracted data from
  - 33% of all pages
  - 41% of machine generated pages
  - 55% of machine generated pages with sufficient html formatting

WhizBang! LABS

# Conclusion

- With only 15% of the work we got 55% of the information that we targeted
- We preserved the meaning of the website data and can put it in a database

WhizBang! LABS

# More to Come?

- Tools developed at WhizBang! Labs, Inc. will significantly improve Global Models, Hand Wrappers, and the ELIJAH approach.
- As the “Spirit of Elijah” spreads throughout the world, technology will assist the massive work.