Automating the Extraction of Genealogical Information from the Web

GeneTIQS

Troy Walker & David W. Embley

Family History Technology Conference
March 25, 2004

Research funded by NSF grant #IIS-0083127

Genealogical Information on the Web

- **Hundreds of thousands of sites**
  - Some professional (Ancestry.com, Familysearch.org)
  - Mostly hobbyist (203,200 indexed by Cyndislist.com)
- **Search engines**
  - “Walker genealogy” on Google: 199,000 results
  - 1 page/minute = 5 months to go through
- **Why not enlist the help of a computer?**

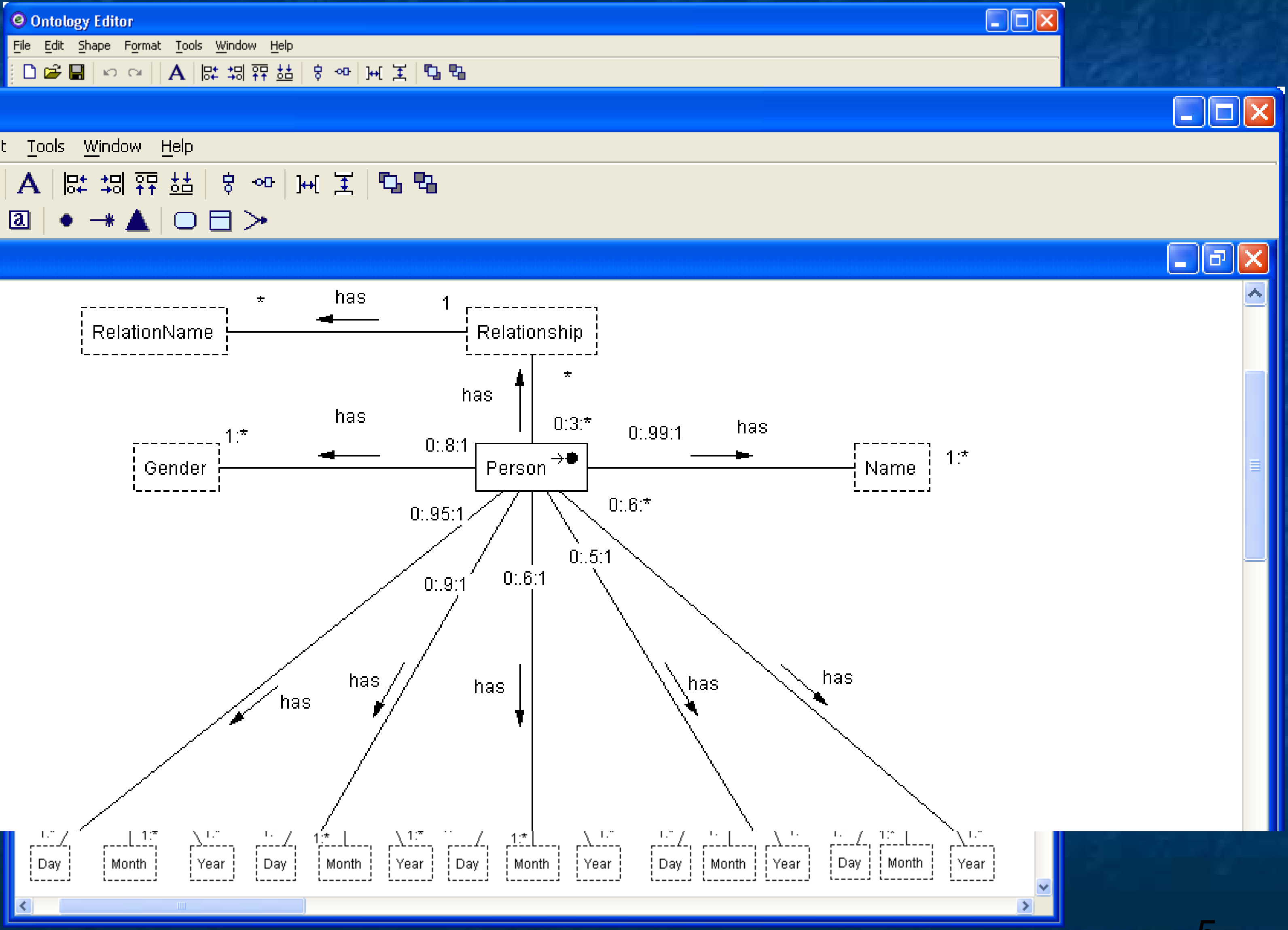
Problems

- No standard way of presenting data
- Sites have differing schemas
- Web pages change
- New pages continuously come on line

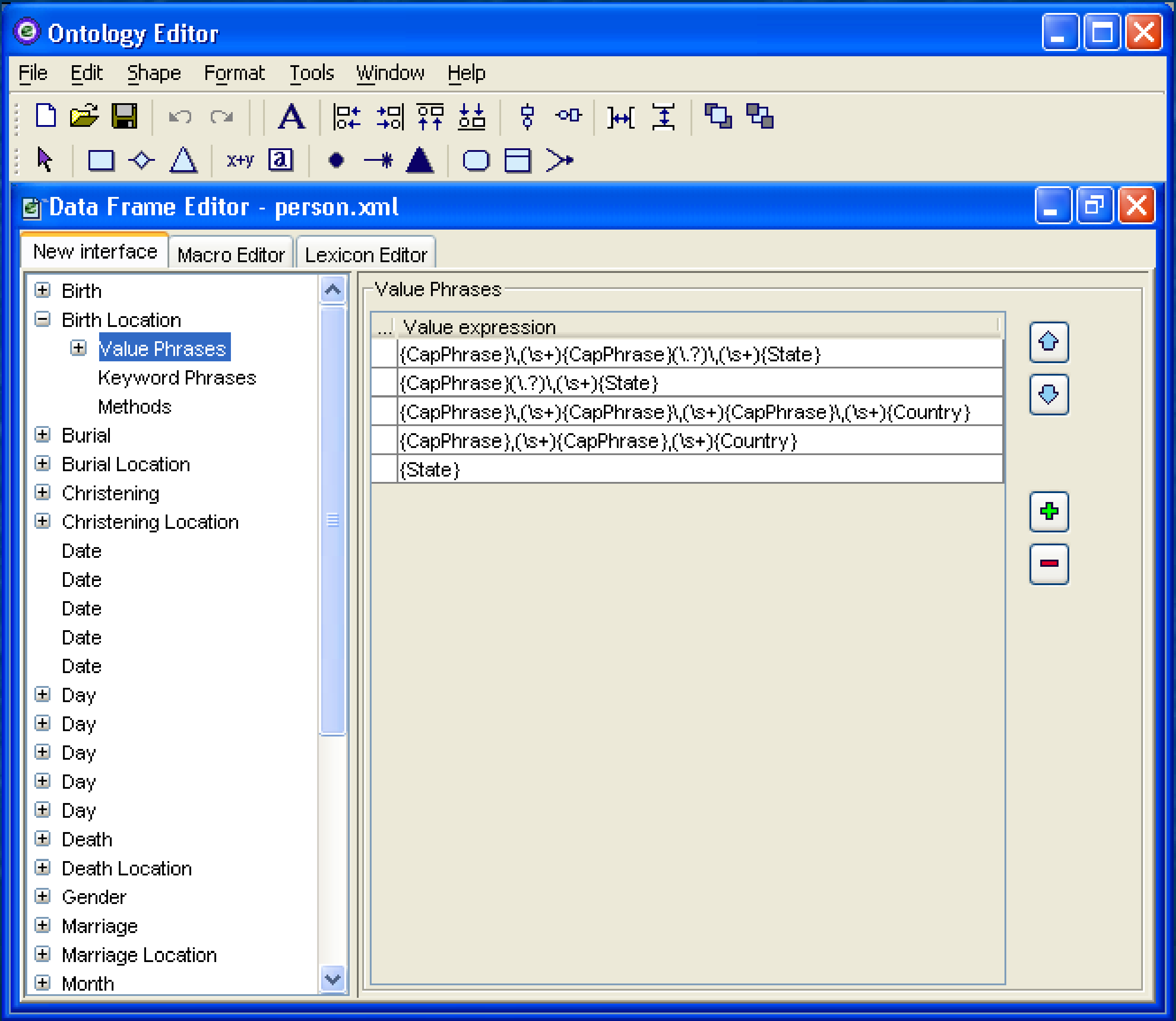
GeneTIQS

- Based on work done by BYU DEG
- Able to extract from:
  - Single-record documents
  - Simple multiple-record documents
  - Complex multiple-record documents
- Robust to changes in pages
- Immediately works for new pages

Person Ontology



Value Matchers



- Birth
- Birth Location
  - Value Phrases
  - Keyword Phrases
  - Methods
- Burial
- Burial Location
- Christening
- Christening Location
- Date
- Date
- Date
- Date
- Date
- Day
- Day
- Day
- Day
- Day
- Death
- Death Location
- Gender
- Marriage
- Marriage Location
- Month

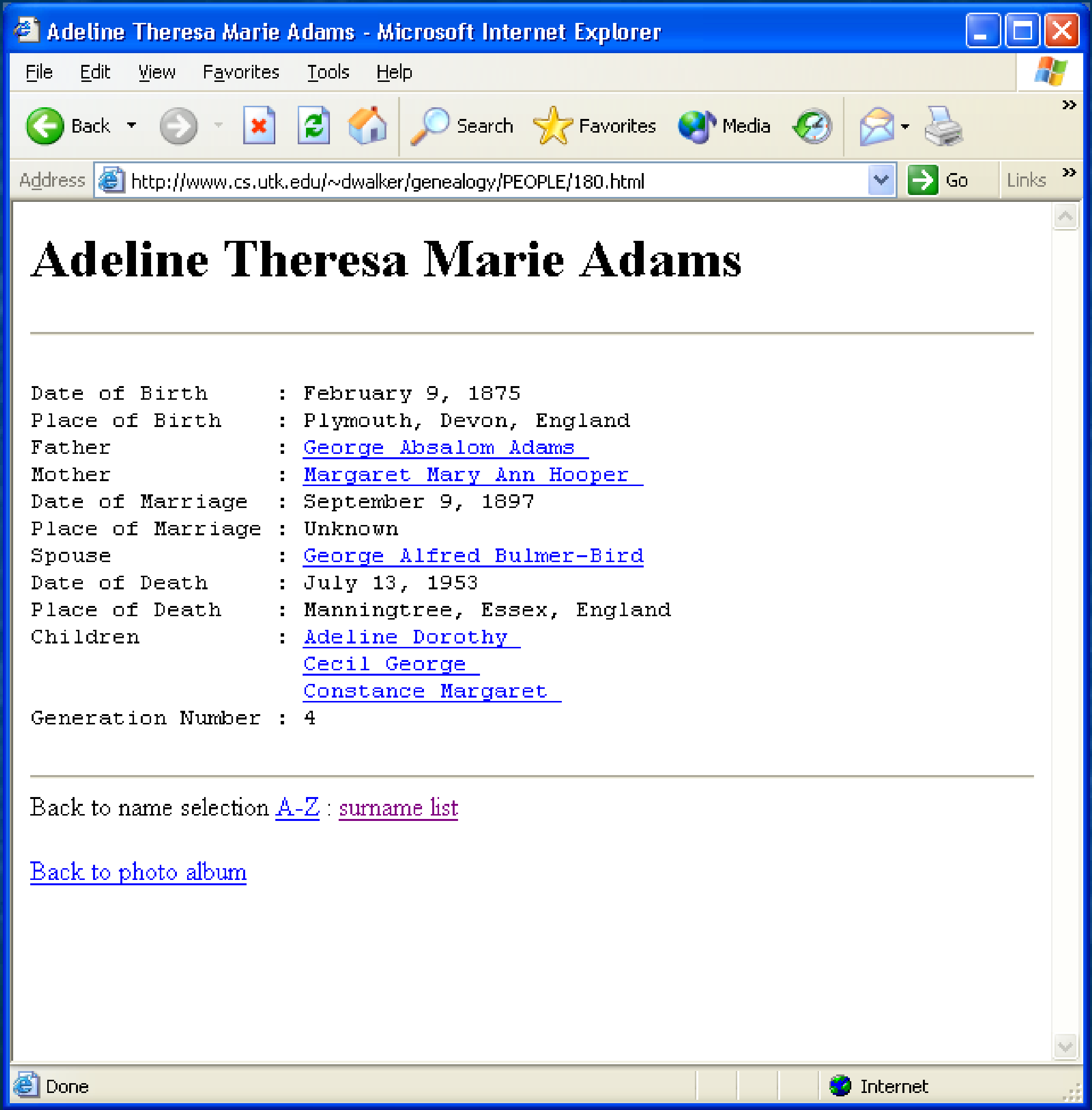
Value Phrases

...	Value expression
	{CapPhrase}\,(\s+){CapPhrase}(\.?)\,(\s+){State}
	{CapPhrase}(\.?)\,(\s+){State}
	{CapPhrase}\,(\s+){CapPhrase}\,(\s+){CapPhrase}\,(\s+){Country}
	{CapPhrase},(\s+){CapPhrase},(\s+){Country}
	{State}

Record Separation

- Separating data related to each person
- Previous technique
  - Combines many heuristics
  - Has problems
    - Assumes multiple records
    - Must be simple separation

Single-Record Document



Adeline Theresa Marie Adams - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://www.cs.utk.edu/~dwalker/genealogy/PEOPLE/180.html
Go
Links
Done
Internet

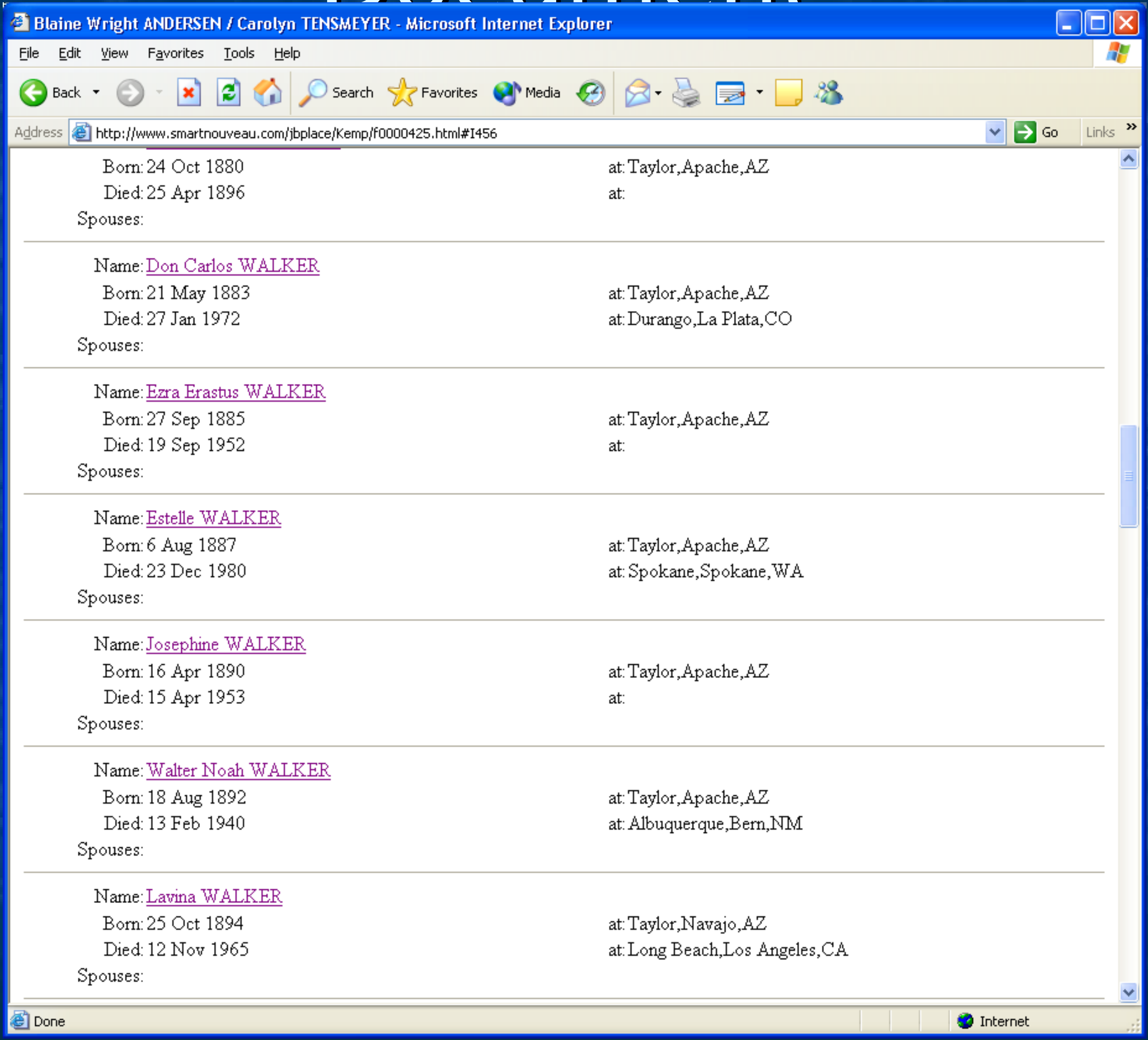
Adeline Theresa Marie Adams

```
Date of Birth      : February 9, 1875
Place of Birth     : Plymouth, Devon, England
Father             : George Absalom Adams
Mother             : Margaret Mary Ann Hooper
Date of Marriage   : September 9, 1897
Place of Marriage  : Unknown
Spouse             : George Alfred Bulmer-Bird
Date of Death      : July 13, 1953
Place of Death     : Manningtree, Essex, England
Children           : Adeline Dorothy
                     Cecil George
                     Constance Margaret
Generation Number  : 4
```

Back to name selection A-Z : surname list

Back to photo album

Simple Multiple-Record Document



Blaine Wright ANDERSEN / Carolyn TENSMEYER - Microsoft Internet Explorer

http://www.smartnouveau.com/jbplace/Kemp/f0000425.html#I456

Born: 24 Oct 1880 at: Taylor,Apache,AZ
Died: 25 Apr 1896 at:
Spouses:

Name: Don Carlos WALKER
Born: 21 May 1883 at: Taylor,Apache,AZ
Died: 27 Jan 1972 at: Durango,La Plata,CO
Spouses:

Name: Ezra Erastus WALKER
Born: 27 Sep 1885 at: Taylor,Apache,AZ
Died: 19 Sep 1952 at:
Spouses:

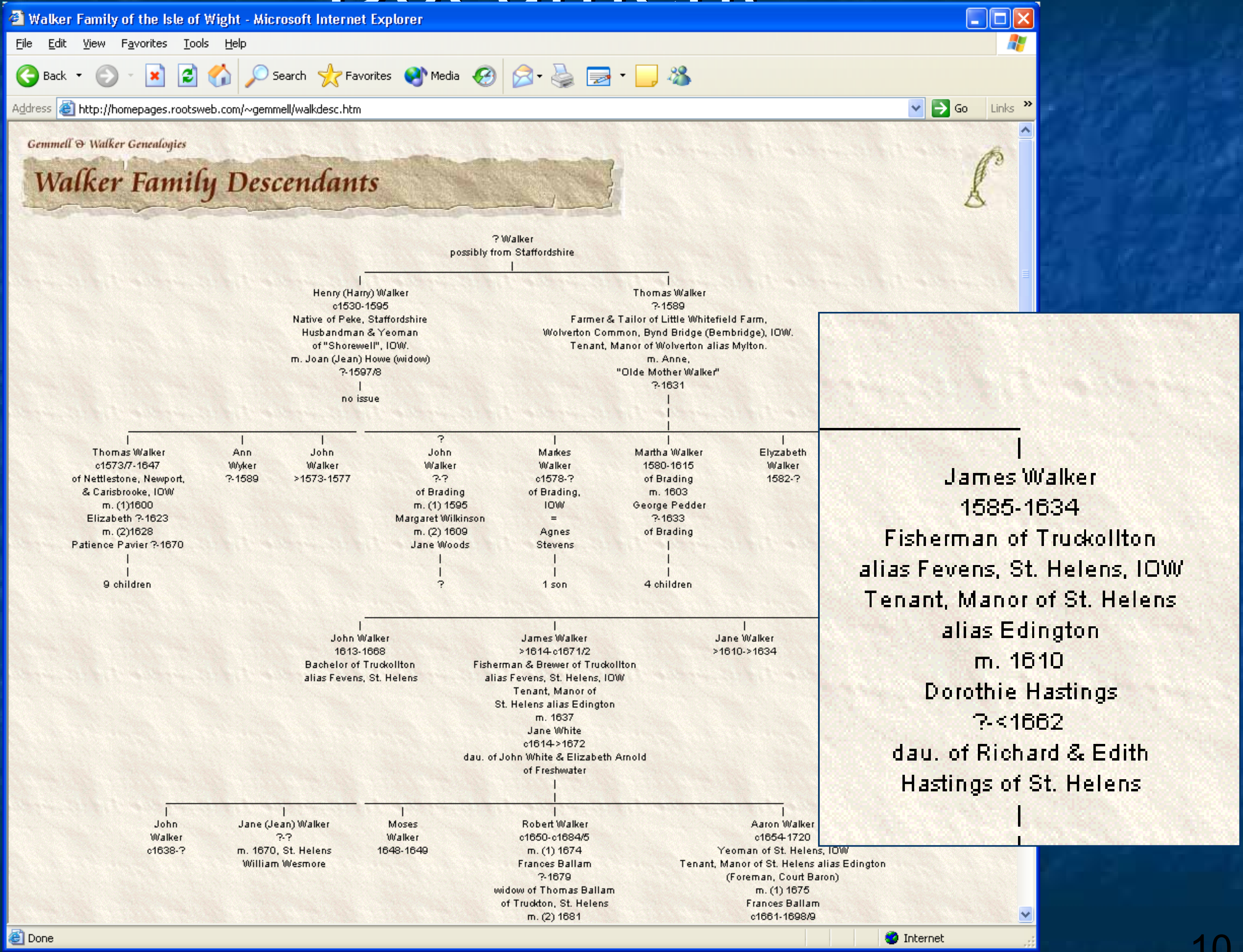
Name: Estelle WALKER
Born: 6 Aug 1887 at: Taylor,Apache,AZ
Died: 23 Dec 1980 at: Spokane,Spokane,WA
Spouses:

Name: Josephine WALKER
Born: 16 Apr 1890 at: Taylor,Apache,AZ
Died: 15 Apr 1953 at:
Spouses:

Name: Walter Noah WALKER
Born: 18 Aug 1892 at: Taylor,Apache,AZ
Died: 13 Feb 1940 at: Albuquerque,Bern,NM
Spouses:

Name: Lavina WALKER
Born: 25 Oct 1894 at: Taylor,Navajo,AZ
Died: 12 Nov 1965 at: Long Beach,Los Angeles,CA
Spouses:

Complex Multiple-Record Document



http://homepages.rootsweb.com/~gemmell/walkdesc.htm

Gemmell & Walker Genealogies

Walker Family Descendants

? Walker
possibly from Staffordshire

Henry (Harry) Walker
c1530-1595
Native of Peke, Staffordshire
Husbandman & Yeoman
of "Shorewell", IOW.
m. Joan (Jean) Howe (widow)
?-1597/8
no issue

Thomas Walker
?-1589
Farmer & Tailor of Little Whitefield Farm,
Wolverton Common, Bynd Bridge (Bembridge), IOW.
Tenant, Manor of Wolverton alias Mylton.
m. Anne,
"Olde Mother Walker"
?-1631

Thomas Walker
c1573/7-1647
of Nettlestone, Newport,
& Carisbrooke, IOW
m. (1)1600
Elizabeth ?-1623
m. (2)1628
Patience Pavier ?-1670
9 children

Ann
Wyker
?-1589

John
Walker
>1573-1577

?
John
Walker
?-?
of Brading
m. (1) 1595
Margaret Wilkinson
m. (2) 1609
Jane Woods
?

Markes
Walker
c1578-?
of Brading,
IOW
=
Agnes
Stevens
1 son

Martha Walker
1580-1615
of Brading
m. 1603
George Pedder
?-1633
of Brading
4 children

Elyzabeth
Walker
1582-?

James Walker
1585-1634
Fisherman of Truckollton
alias Fevens, St. Helens, IOW
Tenant, Manor of St. Helens
alias Edington
m. 1610
Dorothie Hastings
?-<1662
dau. of Richard & Edith
Hastings of St. Helens

John Walker
1613-1668
Bachelor of Truckollton
alias Fevens, St. Helens

James Walker
>1614-c1671/2
Fisherman & Brewer of Truckollton
alias Fevens, St. Helens, IOW
Tenant, Manor of
St. Helens alias Edington
m. 1637
Jane White
c1614->1672
dau. of John White & Elizabeth Arnold
of Freshwater

Jane Walker
>1610->1634

John
Walker
c1638-?

Jane (Jean) Walker
?-?
m. 1670, St. Helens
William Wesmore

Moses
Walker
1648-1649

Robert Walker
c1650-c1684/5
m. (1) 1674
Frances Ballam
?-1679
widow of Thomas Ballam
of Truckton, St. Helens
m. (2) 1681

Aaron Walker
c1654-1720
Yeoman of St. Helens, IOW
Tenant, Manor of St. Helens alias Edington
(Foreman, Court Baron)
m. (1) 1675
Frances Ballam
c1661-1698/9

Vector Space Modeling

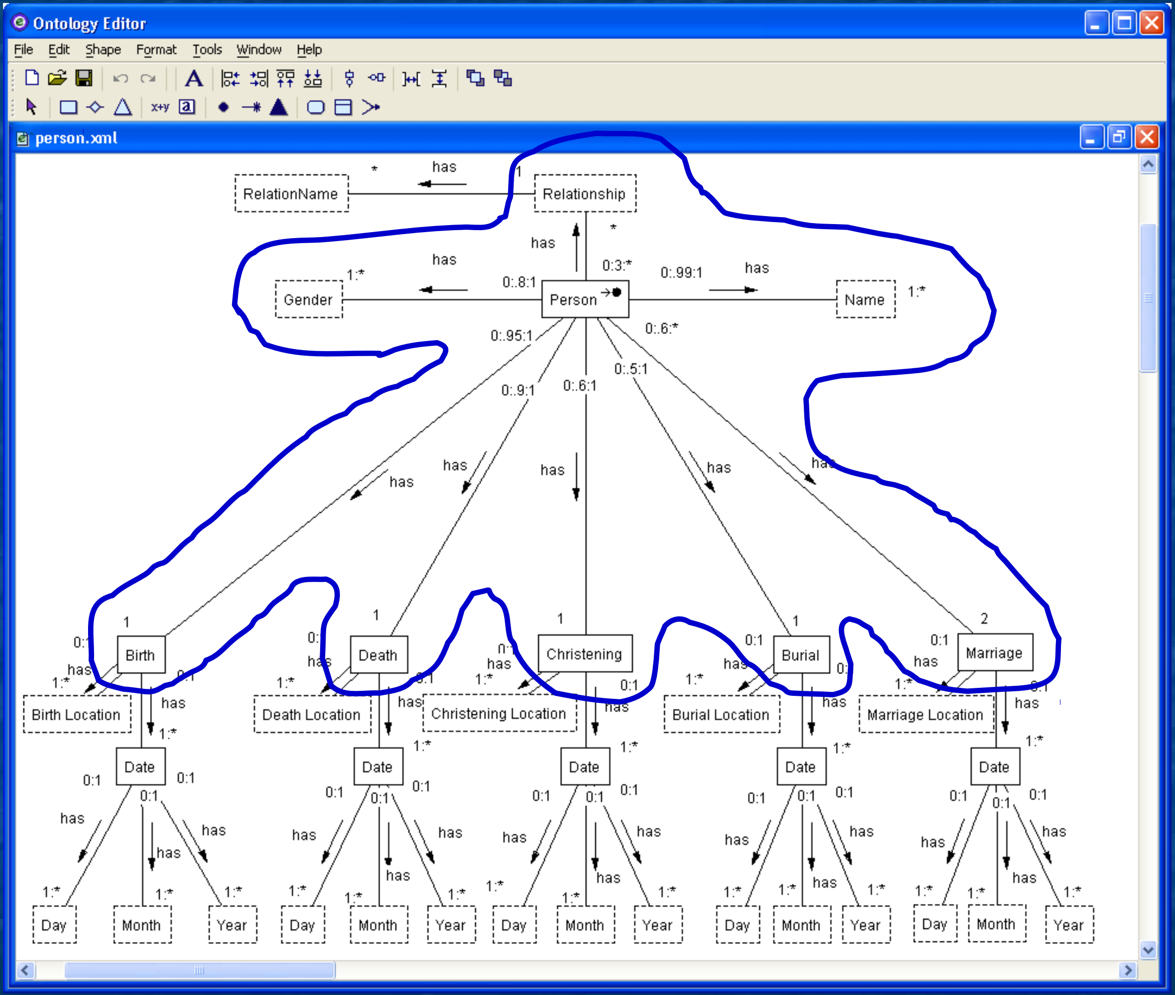
- **Ontology Vector**
- **Compare to candidate records**
  - Cosine measure

    $$v_1 \bullet v_2$$

  - Magnitude measure

    $$|v_1|$$

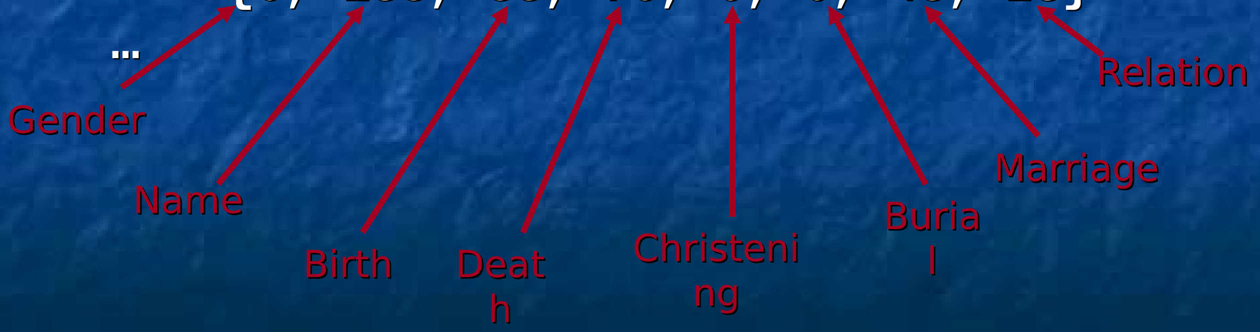
Ontology Vector



{ 0.8, 0.99, 0.95, 0.9, 0.6, 0.5, 0.6, 3.0}

Vector Space Modeling

```
<!DOCTYPE…>           {0, 0, 0, 0, 0, 0, 0, 0}
<html>                {0, 141, 89, 76, 0, 0, 48, 23}
  <head>                {0, 1, 0, 0, 0, 0, 0, 0}
    …
  </head>
  <body>                {0, 140, 89, 76, 0, 0, 48, 23}
   <div>                  {0, 0, 0, 0, 0, 0, 0, 0}
     …header…
   </div>
   <div>                  {0, 138, 88, 76, 0, 0, 48, 23}
     …                …
```



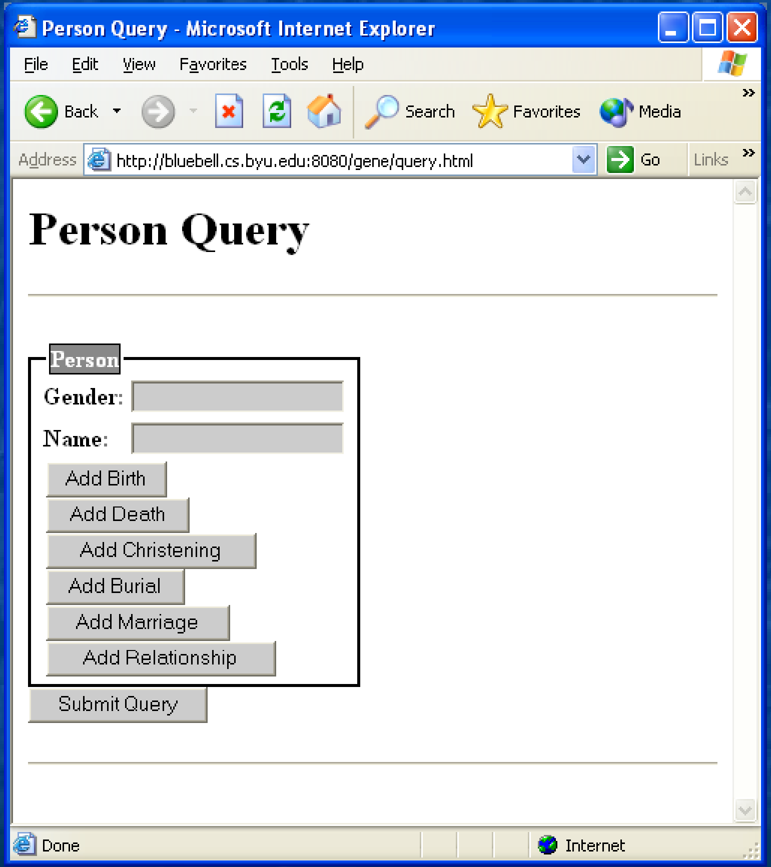
Improvements

- **Differing schemas**
  - Low cosine measures
  - Discarded data
  - Prune dimensions

    {~~0.8~~, 0.99, 0.95, 0.9, ~~0.6~~, ~~0.5~~, 0.6, 3.0}

    {~~0.0~~, 141.0, 89.0, 76.0, ~~0.0~~, ~~0.0~~, 48.0, 23.0}
- **Richness of data in single-record documents**
  - High magnitude measure
  - Higher magnitude to split documents

Demonstration



Person Query - Microsoft Internet Explorer

Address http://bluebell.cs.byu.edu:8080/gene/query.html

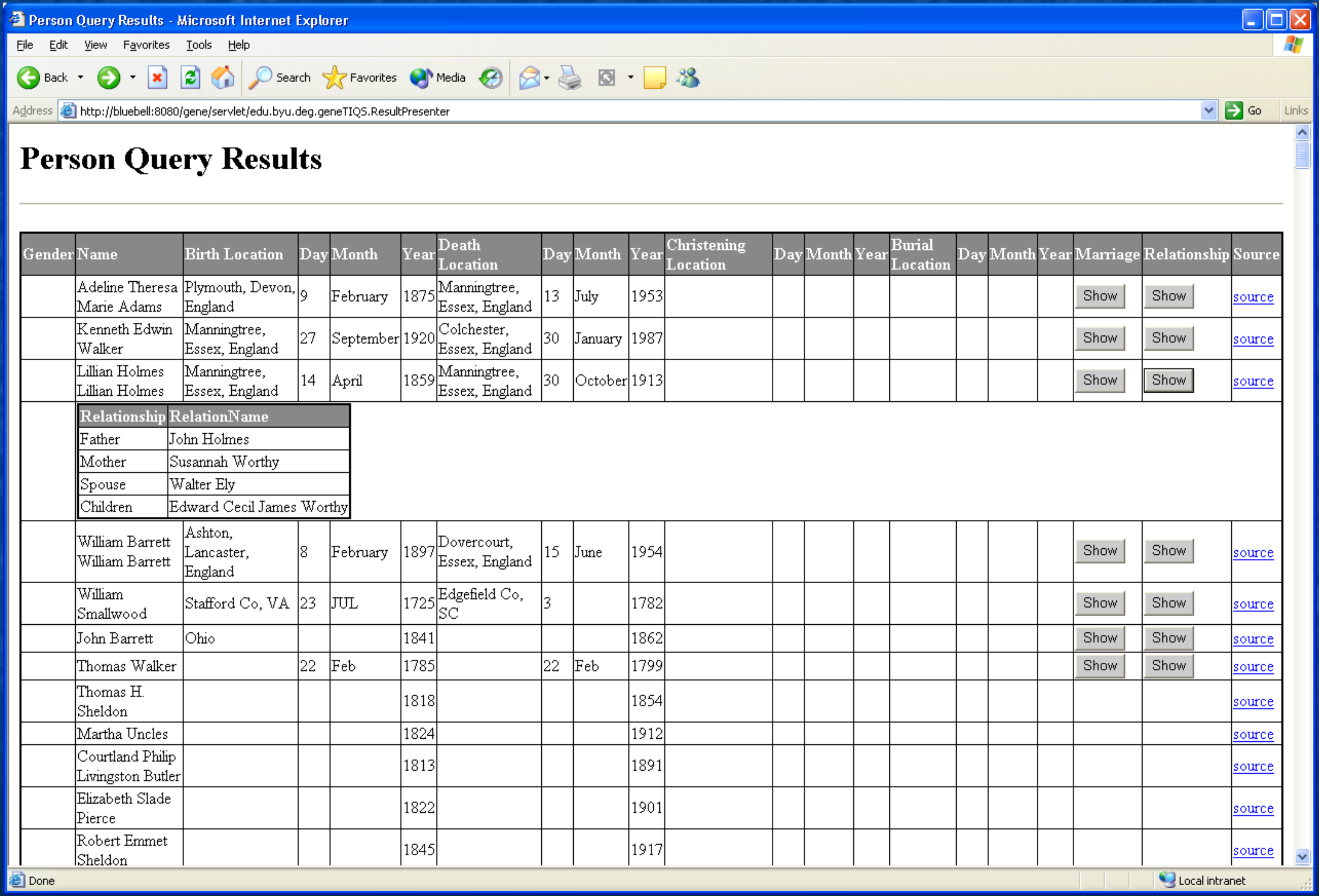
Person Query

Person

Gender:

Name:

Presenting Results



http://bluebell:8080/gene/servlet/edu.byu.deg.geneTIQS.ResultPresenter

Person Query Results

Gender	Name	Birth Location	Day	Month	Year	Death Location	Day	Month	Year	Christening Location	Day	Month	Year	Burial Location	Day	Month	Year	Marriage	Relationship	Source
	Adeline Theresa Marie Adams	Plymouth, Devon, England	9	February	1875	Manningtree, Essex, England	13	July	1953									Show	Show	source
	Kenneth Edwin Walker	Manningtree, Essex, England	27	September	1920	Colchester, Essex, England	30	January	1987									Show	Show	source
	Lillian Holmes Lillian Holmes	Manningtree, Essex, England	14	April	1859	Manningtree, Essex, England	30	October	1913									Show	Show	source

Relationship	RelationName
Father	John Holmes
Mother	Susannah Worthy
Spouse	Walter Ely
Children	Edward Cecil James Worthy

	William Barrett William Barrett	Ashton, Lancaster, England	8	February	1897	Dovercourt, Essex, England	15	June	1954									Show	Show	source
	William Smallwood	Stafford Co, VA	23	JUL	1725	Edgefield Co, SC	3		1782									Show	Show	source
	John Barrett	Ohio			1841				1862									Show	Show	source
	Thomas Walker		22	Feb	1785		22	Feb	1799									Show	Show	source
	Thomas H. Sheldon				1818				1854											source
	Martha Uncles				1824				1912											source
	Courtland Philip Livingston Butler				1813				1891											source
	Elizabeth Slade Pierce				1822				1901											source
	Robert Emmet Sheldon				1845				1917											source

Preliminary Results

- Semi-structured Text
  - 10 single-record documents
  - 3 simple documents containing 268 records
  - 3 complex documents containing 266 records
- Precision and recall for record separation

Record Separation

	Recall	Precision
Single	100%	94.1%
Simple	94.7%	97.3%
Complex	88.3%	93.6%

Conclusion

- Integrate, build on previous DEG work
- Accurate record separation
  - Average recall: 94.3%
  - Average precision: 95.0%
- Ontology based
  - Robust to changes in pages
  - Immediately works with new pages