

# Automatic Extraction From and Reasoning About Genealogical Records: A Prototype

By

Charla J. Woodbury,\* David W. Embley,\* Stephen W.

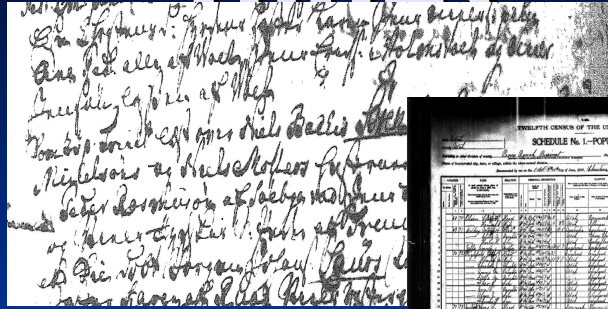
Liddle\*\* \*Department of Computer Science

\*\*Information Systems Department

Brigham Young University

April 28, 2010

# Index

[illegible]

	<b>1. Peter ARSBERG</b>	P
	P - 20. 00. 1972	
	P - 1. 00. 00. 1989	
	P - 1. 00. 00. 1990	
	P - 1. 00. 00. 1991	
	P - 1. 00. 00. 1992	
	P - 1. 00. 00. 1993	
	P - 1. 00. 00. 1994	
	P - 1. 00. 00. 1995	
	P - 1. 00. 00. 1996	
	P - 1. 00. 00. 1997	
	P - 1. 00. 00. 1998	
	P - 1. 00. 00. 1999	
	P - 1. 00. 00. 2000	
	P - 1. 00. 00. 2001	
	P - 1. 00. 00. 2002	
	P - 1. 00. 00. 2003	
	P - 1. 00. 00. 2004	
	P - 1. 00. 00. 2005	
	P - 1. 00. 00. 2006	
	P - 1. 00. 00. 2007	
	P - 1. 00. 00. 2008	
	P - 1. 00. 00. 2009	
	P - 1. 00. 00. 2010	
	P - 1. 00. 00. 2011	
	P - 1. 00. 00. 2012	
	P - 1. 00. 00. 2013	
	P - 1. 00. 00. 2014	
	P - 1. 00. 00. 2015	
	P - 1. 00. 00. 2016	
	P - 1. 00. 00. 2017	
	P - 1. 00. 00. 2018	
	P - 1. 00. 00. 2019	
	P - 1. 00. 00. 2020	
	P - 1. 00. 00. 2021	
	P - 1. 00. 00. 2022	
	P - 1. 00. 00. 2023	
	P - 1. 00. 00. 2024	
	P - 1. 00. 00. 2025	
	P - 1. 00. 00. 2026	
	P - 1. 00. 00. 2027	
	P - 1. 00. 00. 2028	
	P - 1. 00. 00. 2029	
	P - 1. 00. 00. 2030	
	P - 1. 00. 00. 2031	
	P - 1. 00. 00. 2032	
	P - 1. 00. 00. 2033	
	P - 1. 00. 00. 2034	
	P - 1. 00. 00. 2035	
	P - 1. 00. 00. 2036	
	P - 1. 00. 00. 2037	
	P - 1. 00. 00. 2038	
	P - 1. 00. 00. 2039	
	P - 1. 00. 00. 2040	
	P - 1. 00. 00. 2041	
	P - 1. 00. 00. 2042	
	P - 1. 00. 00. 2043	
	P - 1. 00. 00. 2044	
	P - 1. 00. 00. 2045	
	P - 1. 00. 00. 2046	
	P - 1. 00. 00. 2047	
	P - 1. 00. 00. 2048	
	P - 1. 00. 00. 2049	
	P - 1. 00. 00. 2050	
	P - 1. 00. 00. 2051	
	P - 1. 00. 00. 2052	
	P - 1. 00. 00. 2053	
	P - 1. 00. 00. 2054	
	P - 1. 00. 00. 2055	
	P - 1. 00. 00. 2056	
	P - 1. 00. 00. 2057	
	P - 1. 00. 00. 2058	
	P - 1. 00. 00. 2059	
	P - 1. 00. 00. 2060	
	P - 1. 00. 00. 2061	
	P - 1. 00. 00. 2062	
	P - 1. 00. 00. 2063	
	P - 1. 00. 00. 2064	
	P - 1. 00. 00. 2065	
	P - 1. 00. 00. 2066	
	P - 1. 00. 00. 2067	
	P - 1. 00. 00. 2068	
	P - 1. 00. 00. 2069	
	P - 1. 00. 00. 2070	
	P - 1. 00. 00. 2071	
	P - 1. 00. 00. 2072	
	P - 1. 00. 00. 2073	
	P - 1. 00. 00. 2074	
	P - 1. 00. 00. 2075	
	P - 1. 00. 00. 2076	
	P - 1. 00. 00. 2077	
	P - 1. 00. 00. 2078	
	P - 1. 00. 00. 2079	
	P - 1. 00. 00. 2080	
	P - 1. 00. 00. 2081	
	P - 1. 00. 00. 2082	
	P - 1. 00. 00. 2083	
	P - 1. 00. 00. 2084	
	P - 1. 00. 00. 2085	
	P - 1. 00. 00. 2086	
	P - 1. 00. 00. 2087	
	P - 1. 00. 00. 2088	
	P - 1. 00. 00. 2089	
	P - 1. 00. 00. 2090	
	P - 1. 00. 00. 2091	
	P - 1. 00. 00. 2092	
	P - 1. 00. 00. 2093	
	P - 1. 00. 00. 2094	
	P - 1. 00. 00. 2095	
	P - 1. 00. 00. 2096	
	P - 1. 00. 00. 2097	
	P - 1. 00. 00. 2098	
	P - 1. 00. 00. 2099	
	P - 1. 00. 00. 2100	

- # 2

# Problem

- Large amounts of primary genealogical data
- Big projects to index and extract records
- Two independent indexers and adjudication
- Millions of human hours used to index or match records for names and families

# Automated Extraction Solution

- Create a specialized extraction ontology to interpret and label genealogical data
- Add rules and logic that
  - Label family roles - husband, daughter, etc.
  - Link family relationships
    - HUSBAND - WIFE
    - PARENT - CHILD

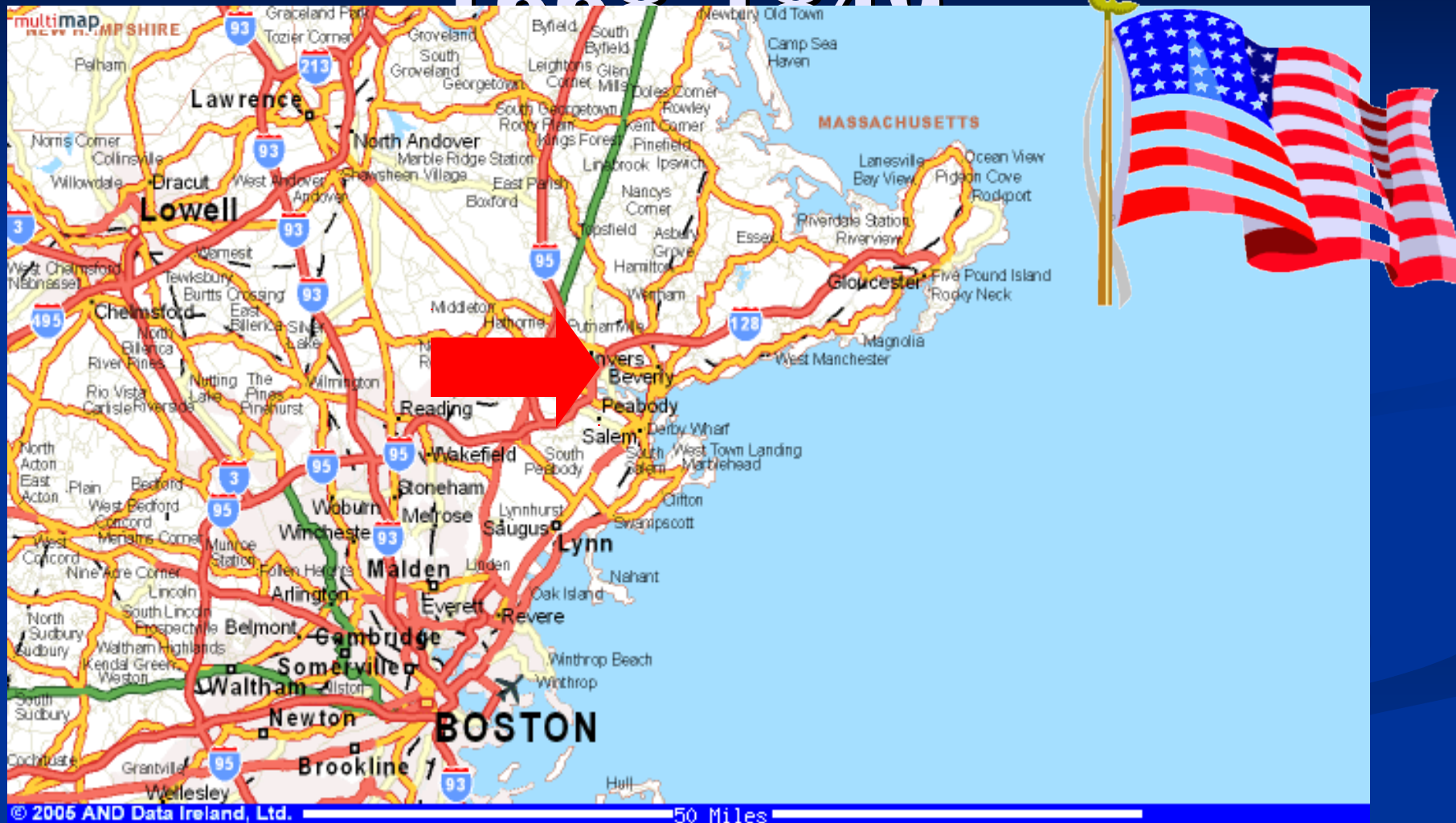
# Outline

1. Data Preparation
2. Ontology Extraction System (OntoES)
3. OWL File and SWRL Rules
4. SPARQL Queries
5. Experimental Results
6. Conclusions

# 1. Data Preparation

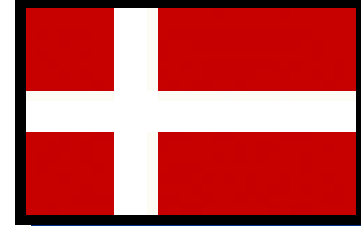
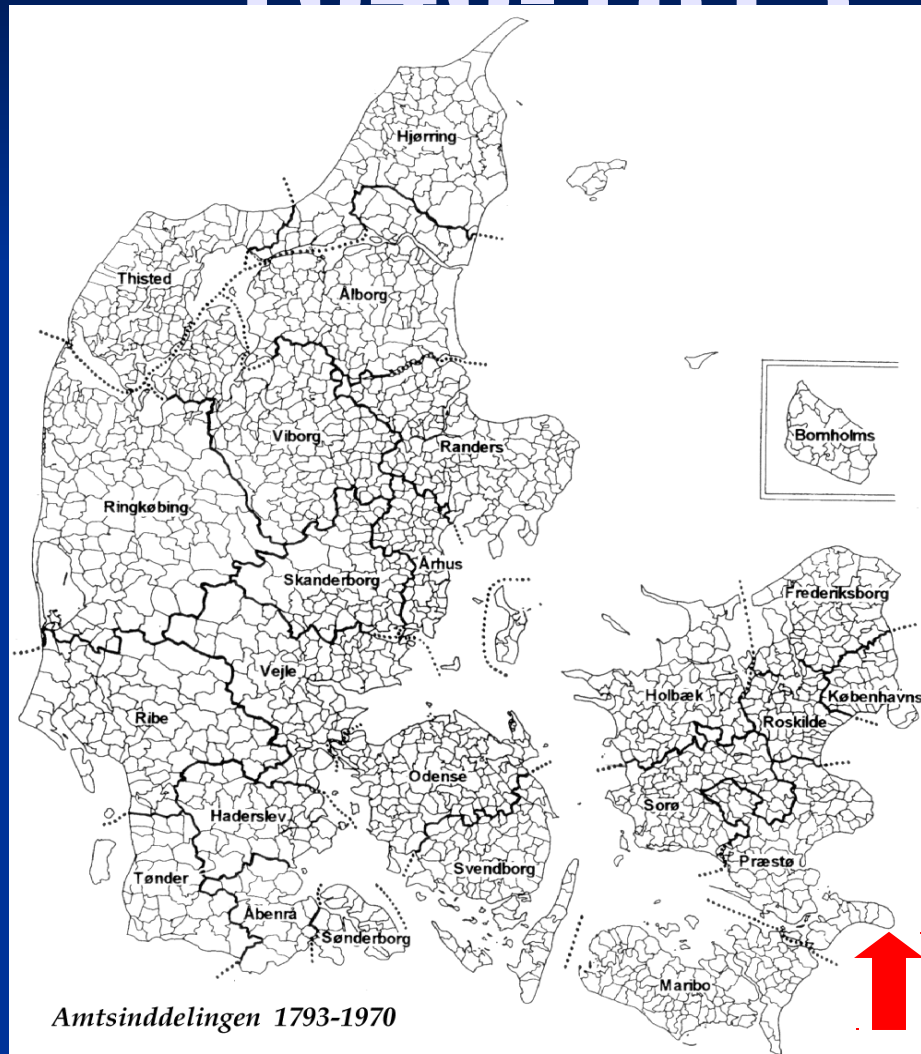
- Collect machine-readable records from three different countries
- Format in HTML format for extraction
- Prepare lexicons for names, places, etc.

# New England Vital Records - Beverly, Massachusetts 1668-1940



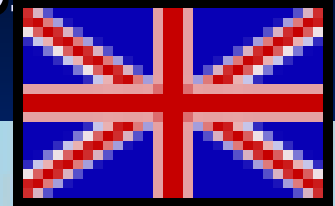


# Danish Parish - Maglebye, Praesto 1646-1813





# English Parish - South Petherton, Somersetshire 1574-1901



# **SOUTH PETHERTON MARRIAGES (from genuki)**

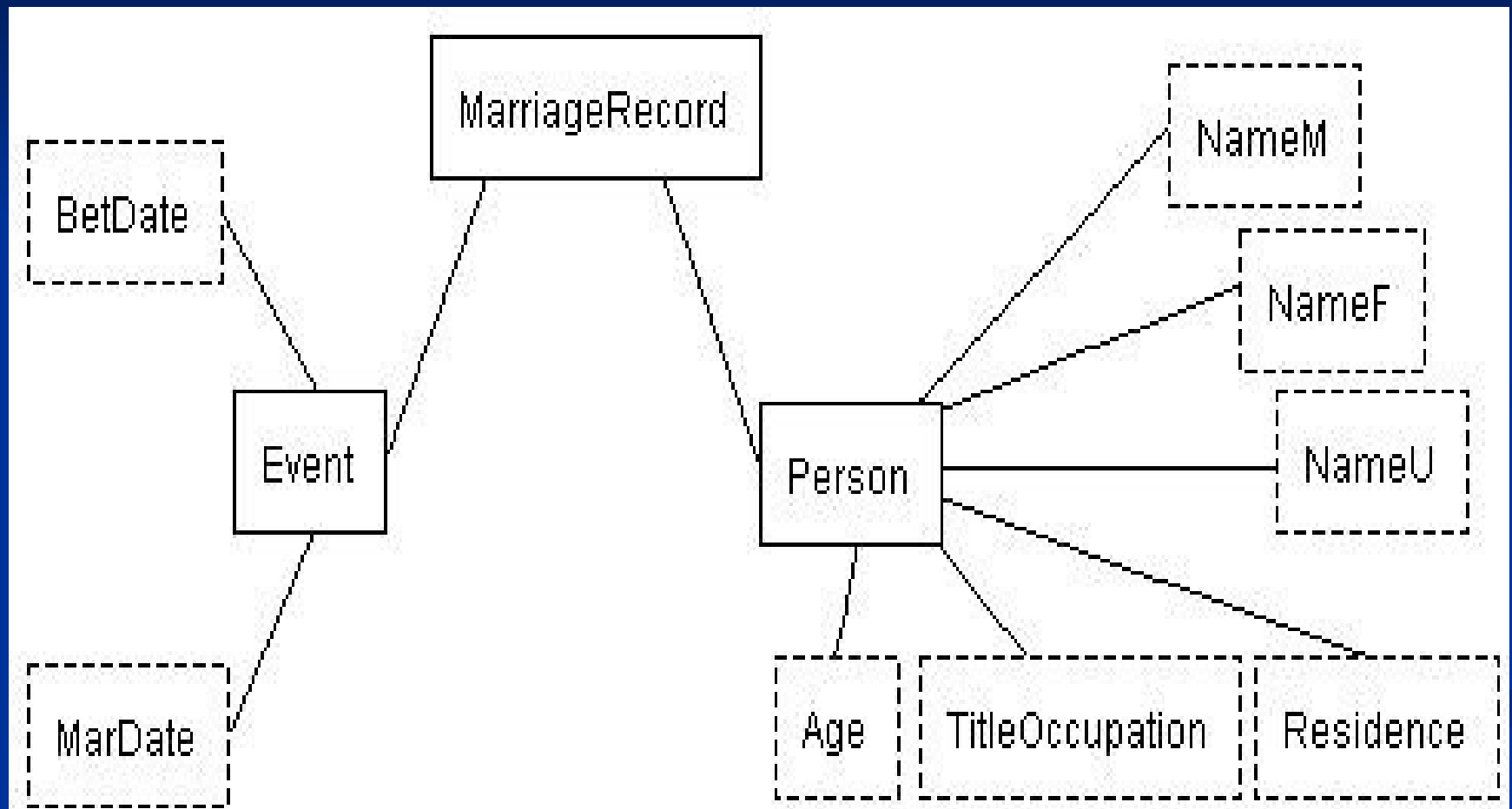
same day 1576 Nicholas Patch and Christian Denman  
26 Jan 1605 Richard Patch and Joan Labor  
25-Sep 1613 John Elliott and Joan Woodbery  
7-Aug 1615 Thomas Prime and Maria Parry  
29-Jan 1616 William Woodbery and Elizabeth Patch  
2-May 1620 William Hillerd and Fortu: Patch  
17-Sep 1622 Nicholas Patch and Elizabeth Owsley  
22-Jan 1627 Richard Patch and Mary White  
15-Jan 1630 Andrew Elliott and Joan Patch  
12-Feb 1639 Andrew Elliott and Joan Pitts

## 2. Ontology Extraction System

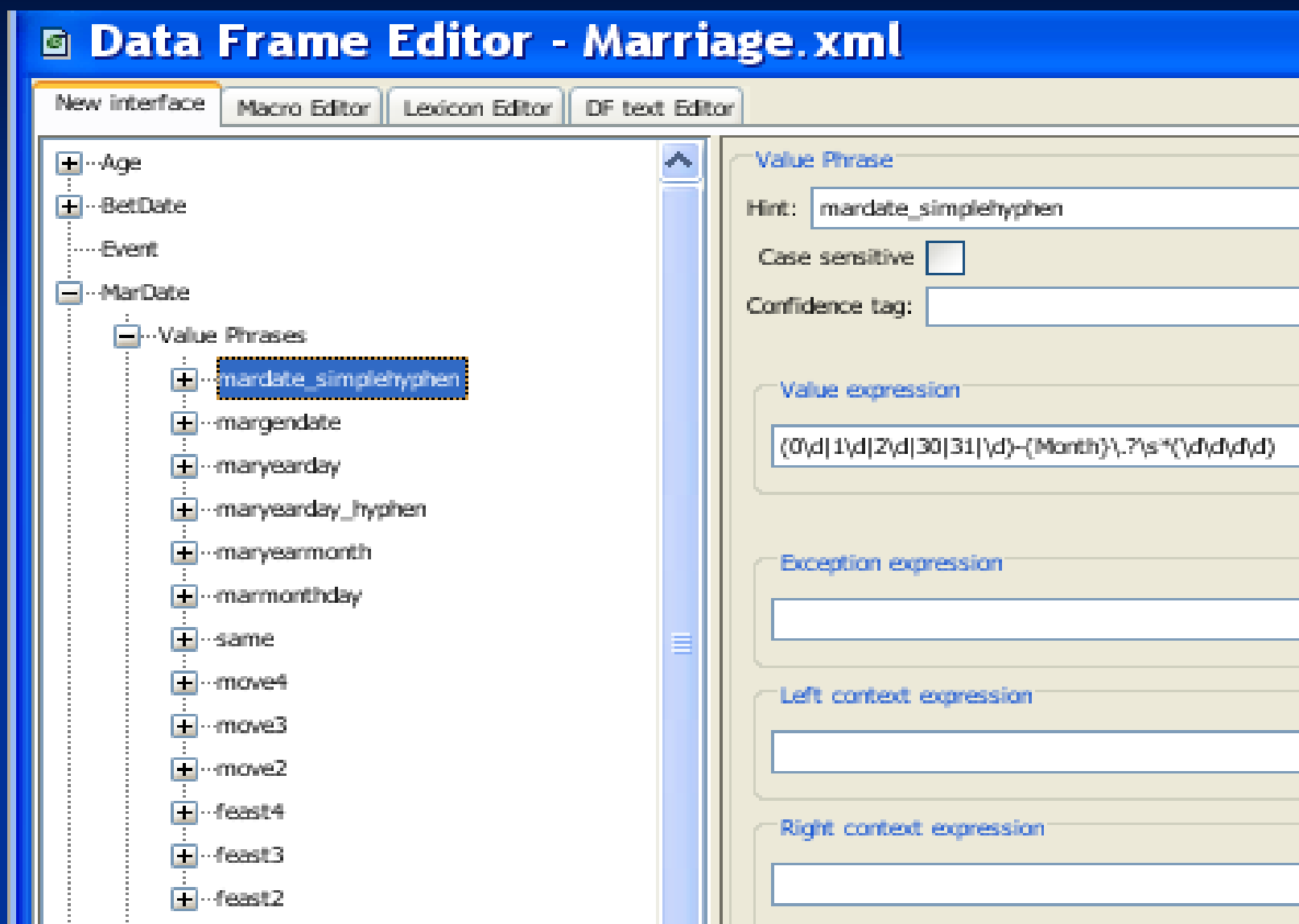
**OntoES**: automatically interpret and correctly label genealogical data using

- Data frames
  - Regular expressions
  - Lexicons
  - Date conversion methods

# Marriage Ontology



# Data Frame Editor



# Sample MONTH LEXICON

- 10ber
- 7ber
- 8ber
- 9ber
- apr
- april
- aprilis
- aug
- august
- augusti
- augustus
- avr
- avril
- avrilis
- dec
- december

- decembr
- decembre
- decembri
- feb
- febr
- februari
- february
- jan
- januarij
- january
- jul
- juli
- julius
- july
- jun
- june

# Object Level

**Data Frame Editor - Marriage.xml**

New interface | Macro Editor | Lexicon Editor | DF text Editor

**Left Panel (Tree View):**

- + Age
- + BetDate
- ... Event
- MarDate (selected)
  - Value Phrases
    - + ..mandate\_simplehyphen
    - + ..margendate
    - + ..maryearday
    - + ..maryearday\_hyphen
    - + ..maryearmonth
    - + ..marmonthday
    - + ..same
    - + ..move4
    - + ..move3
    - + ..move2

**Right Panel (Configuration):**

- ☒ Object set has data frame
- Internal representation**
  - ☐ Object set
    - Object set name: NameU [osmx298]
  - ☒ Data type
    - Type name: string
    - Unit of measure:
- Default canonicalization method**
  - Canonicalization method name: dateToIntYYYYddd
- Output Formatting method**
  - Output format method name: dateOutputDDMMYYYYfromYYYYddd



# CONVERSION METHODS

## inside the ontology

- Regularize date (Julian format: **YYYYddd**)

1620 2-May → **1620093**

- Display stored Julian format as **DD**  
**MMM YYYY**

1620093 → **2 MAY 1620**

16

# Feast Dates

- Fixed Dates

Christmas 1720 → 25 DEC 1720

- Moveable Dates around Easter  
(36 possible Easter dates with leap year variation)

- 1723 Dnica Septuagesima  
1723 → 24 JAN

- Same day as previous entry

# Run Ontology

## ■ Input

- **Ontology** (Created with OntoES)
- **HTML data** (Hypertext Markup Language)

## ■ Output

- **RDF database** (Resource Description Format)
- **OWL file** (Ontology Web Language)

# Ontology Workbench

Ontology Workbench

Ontos Perspective Basic Ontology Editor Perspective

Ontology Editor Marriage.xml

```
graph TD; MarriageRecord --> Event; MarriageRecord --> Person; Event -.-> BetDate; Event -.-> MarDate; Person -.-> NameM; Person -.-> NameF; Person -.-> NameU; Person -.-> Age; Person -.-> TitleOccupation; Person -.-> Residence;
```

The diagram illustrates an ontology for marriage records. It features three main classes: **MarriageRecord**, **Event**, and **Person**. **MarriageRecord** is the root class, which is associated with **Event** and **Person**. **Event** has two attributes: **BetDate** and **MarDate**. **Person** has six attributes: **NameM**, **NameF**, **NameU**, **Age**, **TitleOccupation**, and **Residence**.

South Petherton Marriages

same day 1576 Nicholas Patch and Christian Denman  
26 Jan 1605 Richard Patch and Joan Lavor  
25-Sep 1613 John Elliott and Joan Woodbery  
7-Aug 1615 Thomas Prime and Maria Parry  
29-Jan 1616 William Woodbery and Elizabeth Patch  
2-May 1620 William Hillerd and Fortu: Patch  
17-Sep 1622 Nicholas Patch and Elizabeth Owsley  
22-Jan 1627 Richard Patch and Mary White  
15-Jan 1630 Andrew Elliott and Joan Patch  
12-Feb 1639 Andrew Elliott and Joan Pitts

# Extracted Marriages

Bet Date	MarDate	NameM	NameF	NameU
	same day 1576	Nicholas Patch		Christian Denma n
	26 JAN 1605	Richard Patch	Joan Labor	
	26 SEP 1613	John Elliott	Joan Woodbery	
	7 AUG 1615	Thomas Prime	Maria Parry	
	29 JAN 1616	William Woodbery	Elizabeth Patch	
	2 MAY 1620	William Hillerd		Fortu: Patch
	17 SEP 1622	Nicholas Patch	Elizabeth Owlsey	20 20

# Sample RDF Triples

Person_10	sameAs	Person_10
Person_10	type	Thing
Person_10	type	Person
NameU_0	NameUValue	“Christian Denman”
NameU_0	sameAs	NameU_0
NameU_0	type	Thing
NameU_0	type	NameU
NameM_4	NameMValue	“Nicholas Patch”
NameM_4	sameAs	NameM_4
NameM_4	type	Thing
NameM_4	type	NameM

# OWL File

## ■ OWL HEADER

- `<owl:Class rdf:ID="MarriageRecord"/>`
- `<owl:Class rdf:ID="Person"/>`
- `<owl:Class rdf:ID="NameU"/>`
- `<owl:DatatypeProperty rdf:ID="NameUValue">`
- `<rdfs:domain rdf:resource="#NameU"/>`
- `<rdfs:range rdf:resource="&xsd:string"/>`
- `</owl:DatatypeProperty>`

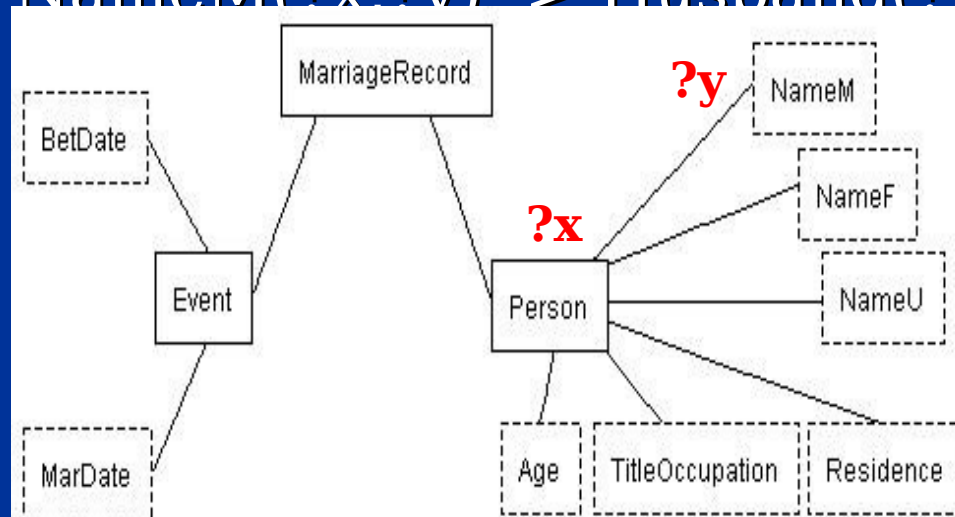
## ■ PERSON - NAMEU

- `<owl:ObjectProperty rdf:ID="Person-NameU">`
- `<rdfs:domain rdf:resource="#Person"/>`
- `<rdfs:range rdf:resource="#NameU"/>`
- `<owl:inverseOf>`
- `<owl:ObjectProperty rdf:ID="NameU-Person"/>`
- `</owl:inverseOf>`
- `</owl:ObjectProperty>`



# 3. OWL File and SWRL Rules

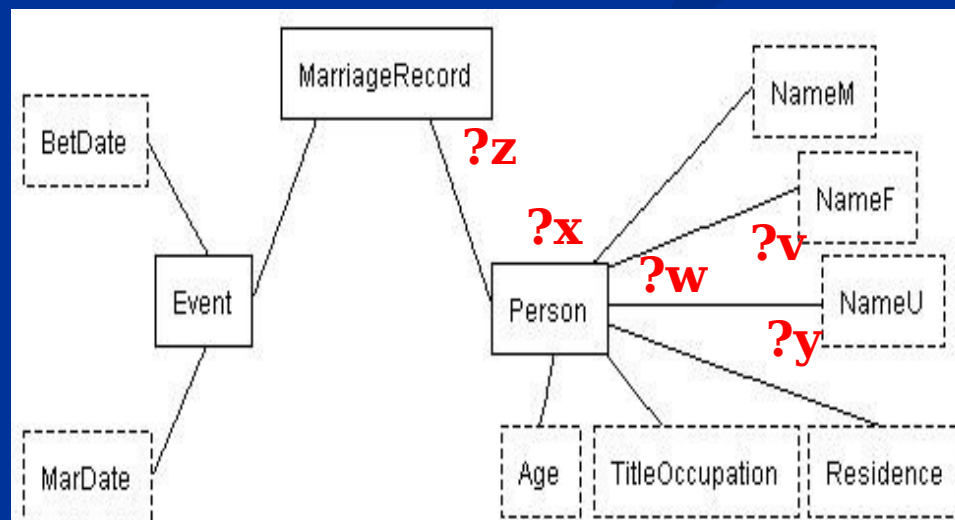
- Define OWL Class
  - Example - Husband
  - `<owl:Class rdf:ID="Husband"/>`
- Define Rule
  - Example - Person with male name is a Husband
  - `Person-NameM(?x ?y) -> Husband(?x)`



# Related Rules

- **NameF is populated then value in NameU is Husband**

$\text{Person-NameU}(\text{?x}, \text{?y}) \wedge \text{Person-NameF}(\text{?w}, \text{?v}) \wedge$   
 $\text{MarriageRecord-Person}(\text{?z}, \text{?x}) \wedge$   
 $\text{MarriageRecord-Person}(\text{?z}, \text{?w})$   
 $\rightarrow \text{Husband}(\text{?x})$



# HusbandOf Rule

$\text{Husband}(\text{?x}) \wedge \text{Wife}(\text{?y}) \wedge \text{MarriageRecord-}$   
 $\text{Person}(\text{?z}, \text{?x})$   
 $\wedge \text{MarriageRecord-Person}(\text{?z}, \text{?y})$   
 $\rightarrow \text{HusbandOf}(\text{?x}, \text{?y})$

# Auxiliary Name Rules

NameM(?x) -> Name(?x)

NameF(?x) -> Name(?x)

NameU(?x) -> Name(?x)

NameMValue(?x) -> NameValue(?x)

NameFValue(?x) -> NameValue(?x)

NameUValue(?x) -> NameValue(?x)

Person-NameM(?x,?y) -> Person-Name(?x,?  
y)

Person-NameF(?x,?y) -> Person-Name(?x,?y)

Person-NameU(?x,?y) -> Person-Name(?x,?y)

# 4. SPARQL Query

## Who is **Husband of** Christian Denman?

PREFIX : <http://www.deg.byu.edu/ontology/Marriage#>

```
SELECT ?Husband
WHERE
{
  ?X :NameValue "Christian Denman" .
  ?Y :Person-Name ?X .
  ?W :HusbandOf ?Y .
  ?W :Person-Name ?V .
  ?V :NameValue ?Husband
}
```

# Query Results

Husband

=====

"Nicholas Patch"^^http://www.w3.org/2001/XMLSchema#string

# Query Results

Husband

=====

"Nicholas Patch"^^http://www.w3.org/2001/XMLSchema#string

## South Petherton Marriages

same day 1576 Nicholas Patch and Christian Denman

26-Jan-1605 Richard Patch and Joan Laver

25-Sep-1611 "Nicholas Patch" because:

7-Aug-1615 NameValue("Nicholas Patch") and Name-

29-Jan-1616 NameValue(n1, "Nicholas Patch")

2-May-1620 and Name(n1) is NameM(n1) and Person-

17-Sep-1621 NameM(p1, n1)

22-Jan-1627 NameValue("Christian Denman") and Name-

15-Jan-1630 NameValue(n2, "Christian Denman")

12-Feb-1631 and Name(n2) is NameU(n2) and Person-

NameU(p2, n2)

Husband(p1) because:

Person-NameM(p1, n1)

Wife(p2) because:

Person-NameU(p2, n2) and Person-MarriageRecord(p2, r1)

and MarriageRecord-Person(r1, p1) and Person-NameM(p1, n1)



# 5. Experimental Results

- Extraction Results
- American Extraction Problem
- Rule Results

# Extraction Results

	MARRIAGES	ENTITIES	RECALL	%	ERRORS	PRECISION
English	188	594	588	99.0%	8	98.7%
American	608	1824	1630	89.4%	34	98.0%
Danish	171	543	538	99.1%	10	98.2%
	BIRTHS					
English	3153	9489	9394	99.0%	61	99.4%
American	675	2055	1809	88.0%	33	98.2%
Danish	677	2061	2042	99.1%	15	99.3%
	DEATHS					
English	3458	8675	8589	99.0%	86	99.0%

# American Difficulty

## BIRTH

WOODBURY, Charles Henry [Charles William, P. R. 4.], s. Henry [housewright. dup.] and Henrietta (Galloup), Dec. 4, 1845.

- Extra information inside brackets & parentheses
  - Charles William – twin of Charles Henry
  - Henry [housewright] – identified as NAME
  - Henrietta (Galloup) – identified as NAME

# Rules Results

- 100% Precision and Recall  
(Once rules are well-defined, the results are perfect.)
- Database Size  
(The RDF database 100x larger when rule triples are added.)
  - NEW PROPERTIES – husband, wife, parent, child
  - NEW LINKS

# 6. Conclusions

- Speed up data indexing
- Make production of a full index easier
- Ground the index in original documents
- Provide for inferred facts
- Simplify as well as augment record search
- Help link records and form family groups and ancestral lines