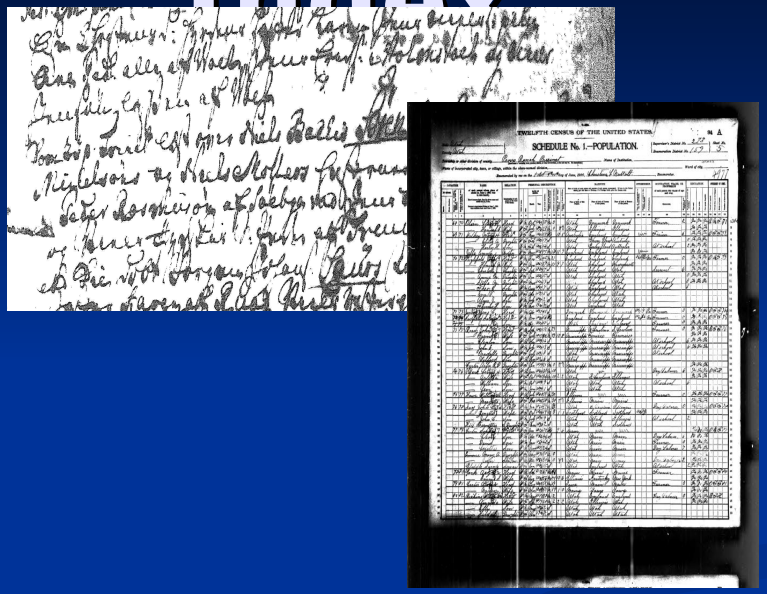
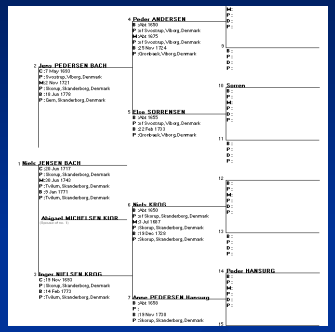
Automatic Extraction From and Reasoning About Genealogical Records: A Prototype

By
Charla J. Woodbury,* David W. Embley,* Stephen W. Liddle** *Department of Computer Science
**Information Systems Department
Brigham Young University
April 28, 2010

Digital Images – Human Index





- Large number of competing family history websites
  - Digital images
  - Human indexes
- Researchers hunting through records and indexes to put families together

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

4

Outline

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

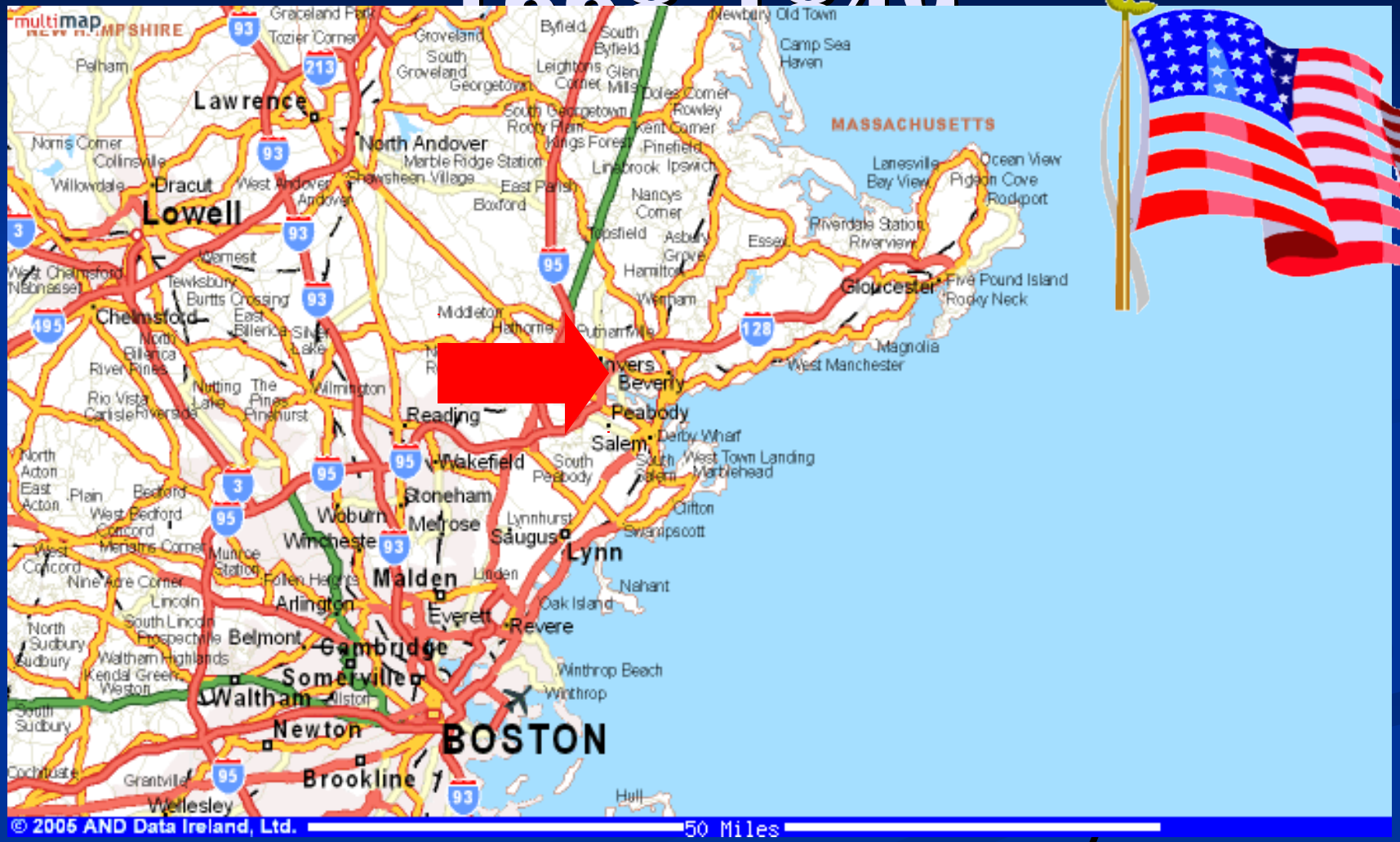
5

1. Data Preparation

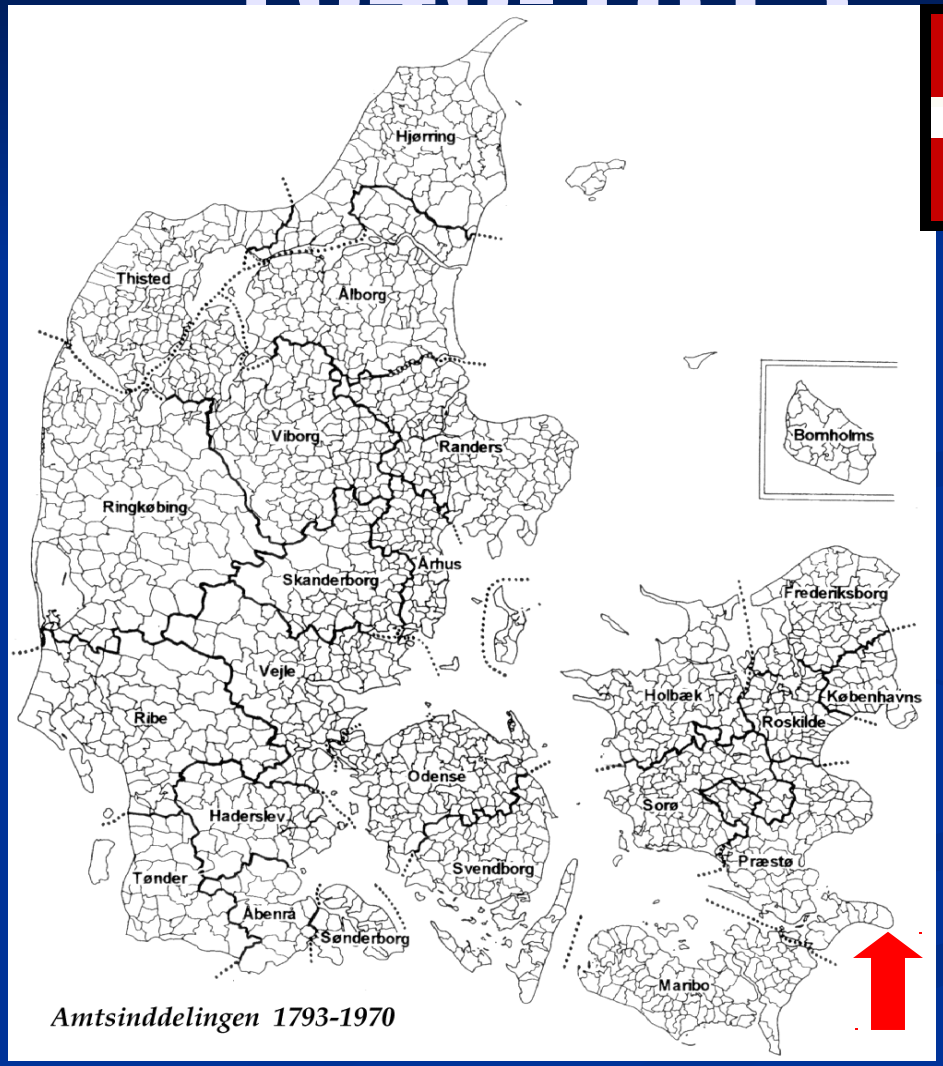
- Collect machine-readable records from three different countries
- Format in HTML format for extraction
- Prepare lexicons for names, places, etc.

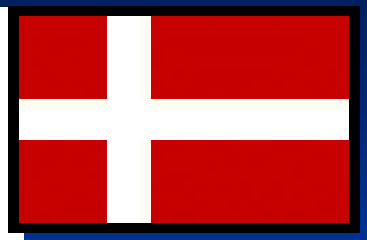
6

New England Vital Records – Beverly, Massachusetts 1668-1840



Danish Parish – Maglebye, Praesto 1646-1813





8

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 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

10

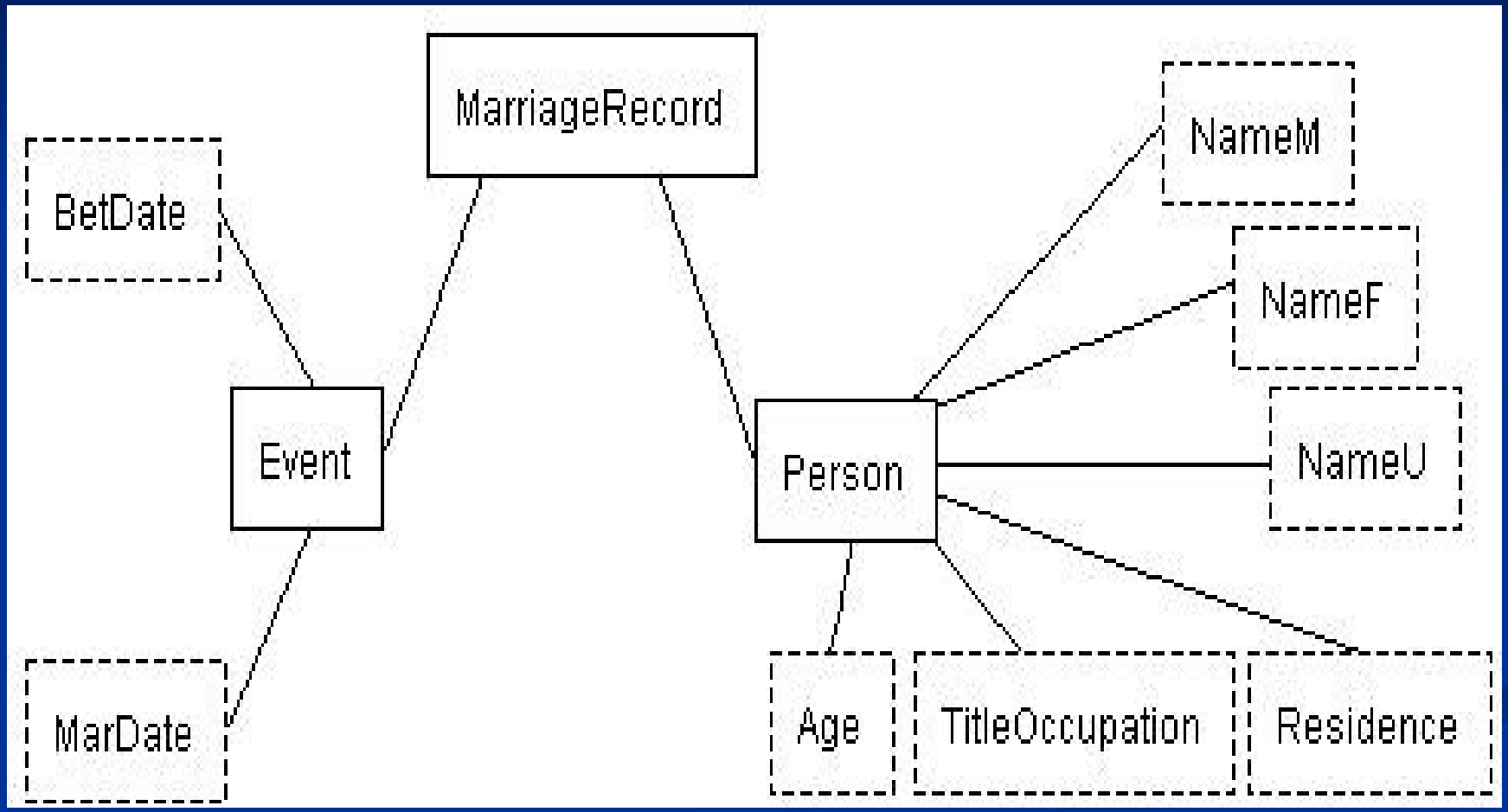
2. Ontology Extraction System

OntoES: automatically interpret and correctly label genealogical data using

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

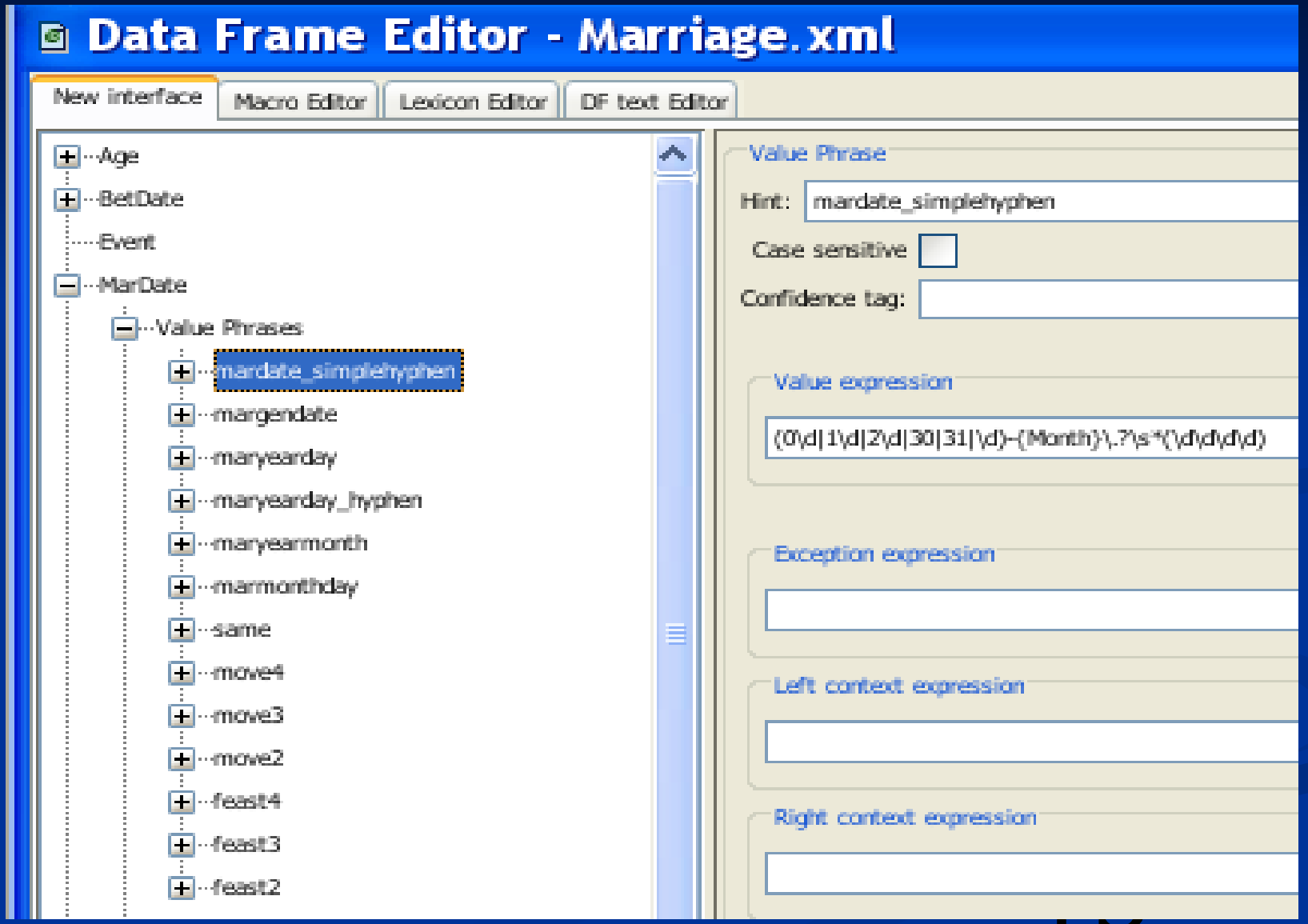
11

Marriage Ontology



12

Data Frame Editor

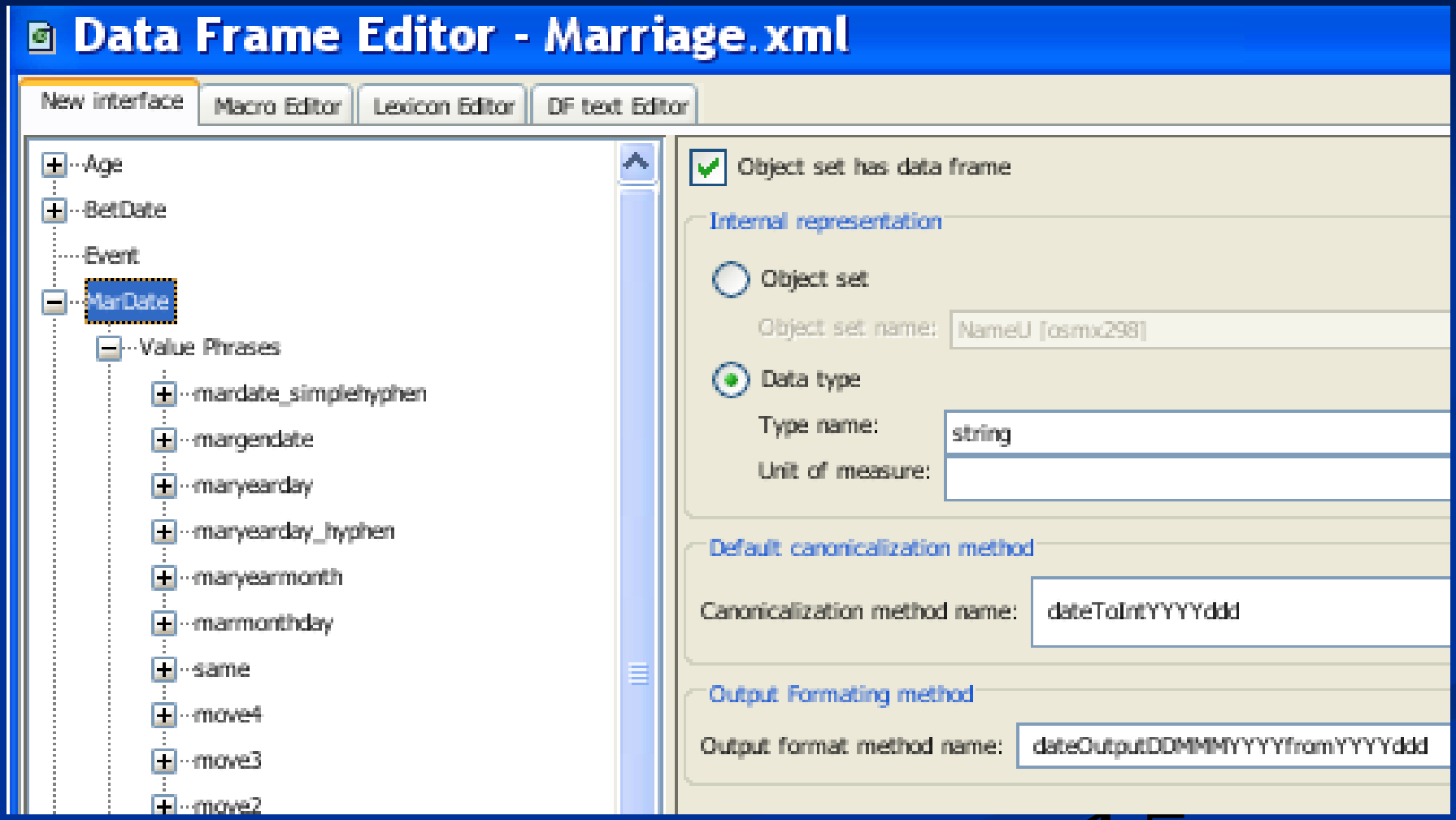


Sample MONTH LEXICON

- 1Ober
- 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

14

Object Level



15

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 → 24 JAN 1723

- Same day as previous entry

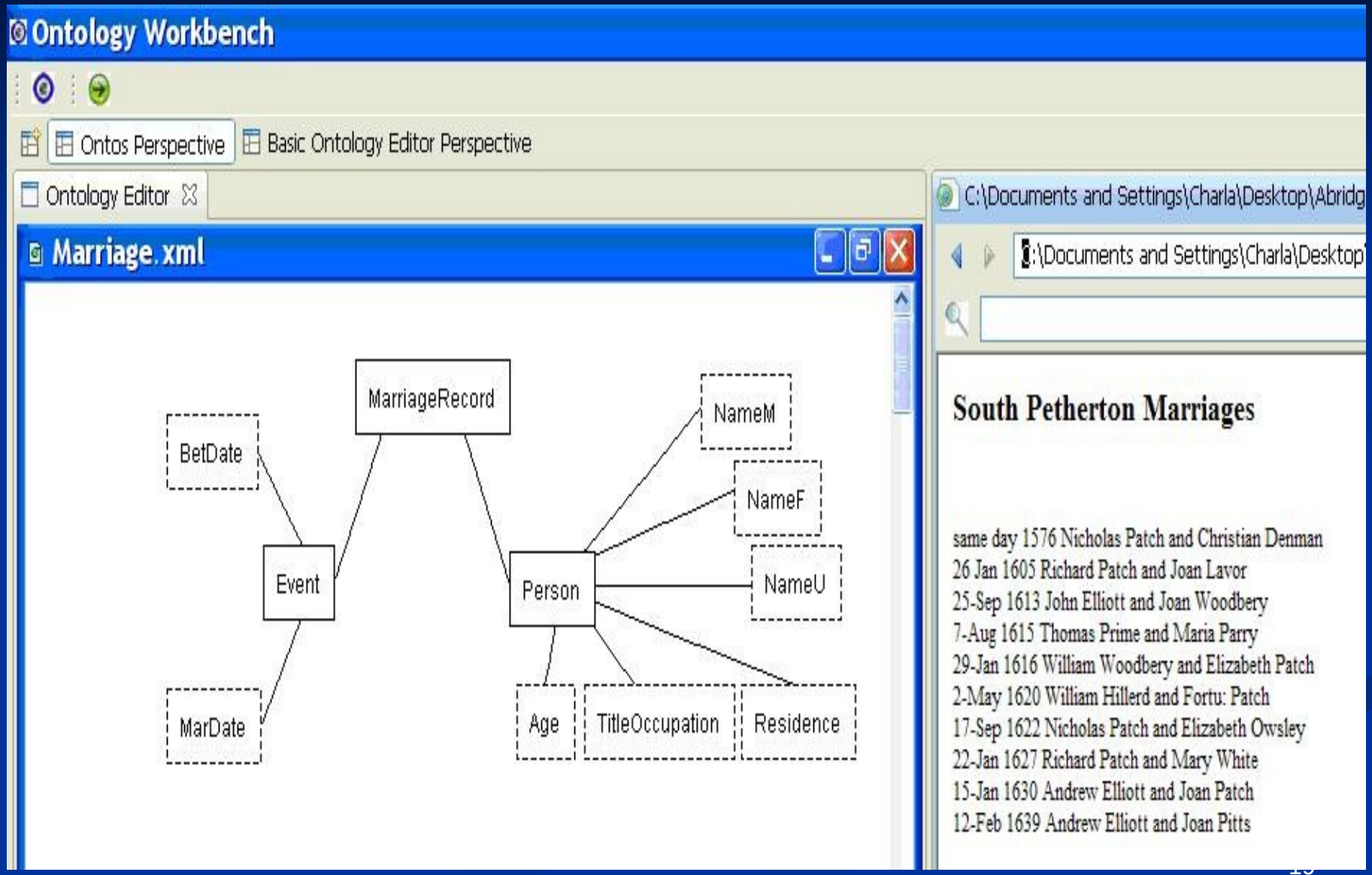
17

Run Ontology

- Input
  - Ontology (Created with OntoES)
  - HTML data (Hypertext Markup Language)
- Output
  - RDF database (Resource Description Format)
  - OWL file (Ontology Web Language)

18

Ontology Workbench



Extracted Marriages

Bet Date	MarDate	NameM	NameF	NameU
	same day 1576	Nicholas Patch		Christian Denman
	26 JAN 1605	Richard Patch	Joan Lavor	
	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	

Sample RDF Triples

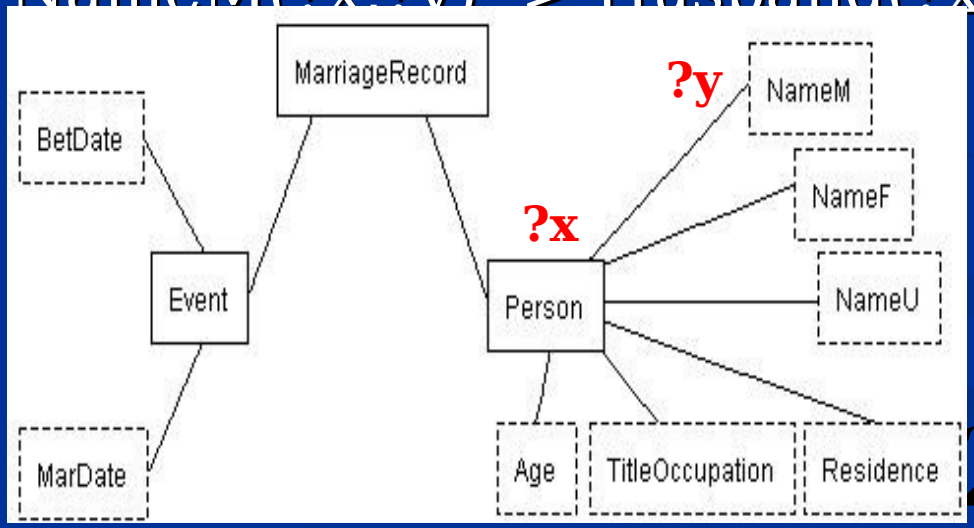
Person_10	sameAs	Person_10
Person_10	type	Thing
Person_10	type	Person
NameU_0	NameUValue	“Christian Denman”
NameU_0	sameAs	NameU_0
NameU_0	type	Thing
NameU_0	type	NameU
NameM_4	NameMValue	“Nicholas Patch”
NameM_4	sameAs	NameM_4
NameM_4	type	Thing
NameM_4	type	NameM

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)



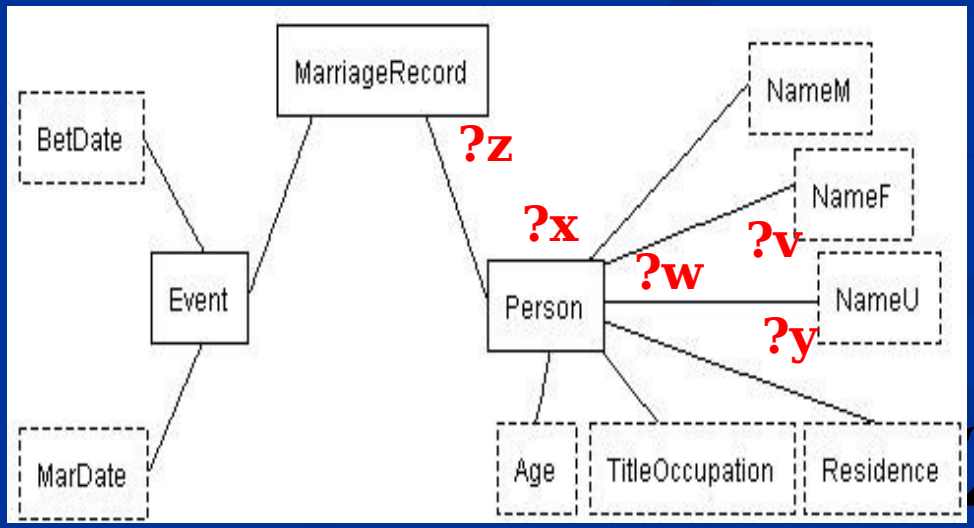
23

Related Rules

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

  Person-NameU(?x,?y) ^ Person-NameF(?w,?v) ^
  MarriageRecord-Person(?z,?x) ^
  MarriageRecord-Person(?z,?w)

  -> Husband(?x)



24

HusbandOf Rule

Husband(?x) ^ Wife(?y) ^ MarriageRecord-Person(?z,?x)

^ MarriageRecord-Person(?z,?y)

-> HusbandOf(?x,?y)

25

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)

26

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
}
```

27

Query Results

Husband

==========

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

28

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 Lavor

25-Sep 161

7-Aug 1615

29-Jan 161

2-May 162

17-Sep 162

22-Jan 162

15-Jan 163

12-Feb 163

"Nicholas Patch" because:
NameValue("Nicholas Patch") and NameNameValue(n1, "Nicholas Patch")
and Name(n1) is NameM(n1) and PersonNameM(p1, n1)
NameValue("Christian Denman") and NameNameValue(n2, "Christian Denman")
and Name(n2) is NameU(n2) and PersonNameU(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 PersonNameM(p1, n1)

5. Experimental Results

- Extraction Results
- American Extraction Problem
- Rule Results

30

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 %	83	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

32

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

33

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

34