Family History Research on the Semantic Web: Building a Semantic Prototype for Danish Research

Charla Woodbury and David W. Embley
BYU Computer Science Department
Email: charlajw@cs.byu.edu embley@cs.byu.edu

Introduction

One of the most popular pursuits on the Internet is genealogy or family history research. The Internet is perfect for sharing family information and for publishing completed research. Popular genealogical research sites have some of the highest webpage hit statistics recorded. Even though the Ellis Island immigration website anticipated a large volume of traffic as it unveiled its new website which included a large database of immigrants into the port of New York City, the system was brought to its knees within minutes by enthusiastic family researchers. It took more than three months for the improvements needed to allow the website to stay online.

However, there are underlying problems as there are in other areas of Internet research that have not yet been addressed. Most genealogical researchers are not well trained in research principles or in the fine art of Internet research, nor are they aware of the vast amounts of primary data now being made available from the libraries and archives of the world or how to get to them. What is needed is a smarter way to research.

At the same time, the next version of the Internet will quite probably be a semantic web – a web that will have an understanding of how words are defined in context. Imagine an Internet that can distinguish between using the word '**DATE**' to mean:

- An indication of time giving the day, month, and year
- The verb-form which is the act of assigning a date to an object or artifact
- A fruit that we eat
- A romantic experience
- The verb-form which is the act of going on a romantic experience with someone

Actually **DATE** is also a **NAME** which has the potential of being extremely confusing in the world of family history research. Mr. C.J. Date is well known along with Edgar F. Codd as the Codd and Date famous for defining the different levels of normal forms in database design. The important point is that the semantic web would be able to allow a machine to differentiate when '**DATE**' was a name and when it was a calendar date.

The semantic web is moving closer to being a reality, but no real application with real data has been developed in the genealogical arena that would show

how this development would impact research the Internet. There are several real problems that relate to many aspects of web research that will addressed by this research.

The purpose of this research is then to:

1) Design a prototype of family history research for a small geographical area for the semantic web using real research data. The geographical area proposed is Nim District, Skanderborg, Denmark.
2) Test and evaluate the prototype for accuracy in terms of precision and recall, for ease of use, for speed of delivery, and for the clarity of format.

Prototype

In order to build a prototype, two structures need to be built and several problems solved:

1. An Ontology – semantic model for family history (BYU Ontos)
2. Several selected annotated web pages (Web Ontology Language OWL proposed IEEE Feb 2004 will be used to annotate.)
3. Solutions for special genealogical problems

1. Ontologies

An ontology consists of both high-level and low-level descriptions of what entities there are and how they are related.

The Data Extraction Group (DEG) at Brigham Young University has previously developed a tool called Ontos that creates ontologies using modeling techniques. This tool can be used to find, label, and then search web pages. In the future semantic web, such an ontology would be attached to every web page after it was used to automatically mark a designated web page with semantic mark-ups.

Now consider a person who decides to do family history research for the first time on their Danish family lines.

- Where do they go?
- What records do they look for?
- How do they handle records in Danish?
- How can they tell when the records match their family?

With a Danish genealogical ontology on the semantic web in a Danish research portal, these questions would be answered. Whether or not these records matched the search target family is the exception. Even then, the ontology would offer a great deal of help and guidance in evaluating how well the records match.

2. Web Pages with Semantic Annotations

Semantic annotations are extra mark-ups that add semantic meaning. The World Wide Web Consortium (W3C)[1] has proposed a new mark-up language for this purpose in February 2004 called OWL (Web Ontology Language). OWL mark-ups when combined with the built-in web page ontology can give semantic meaning to any web page, and yet they can be quite invisible.

[1] [See http://www.w3.org/TR/2004/REC-owl-features-20040210/]

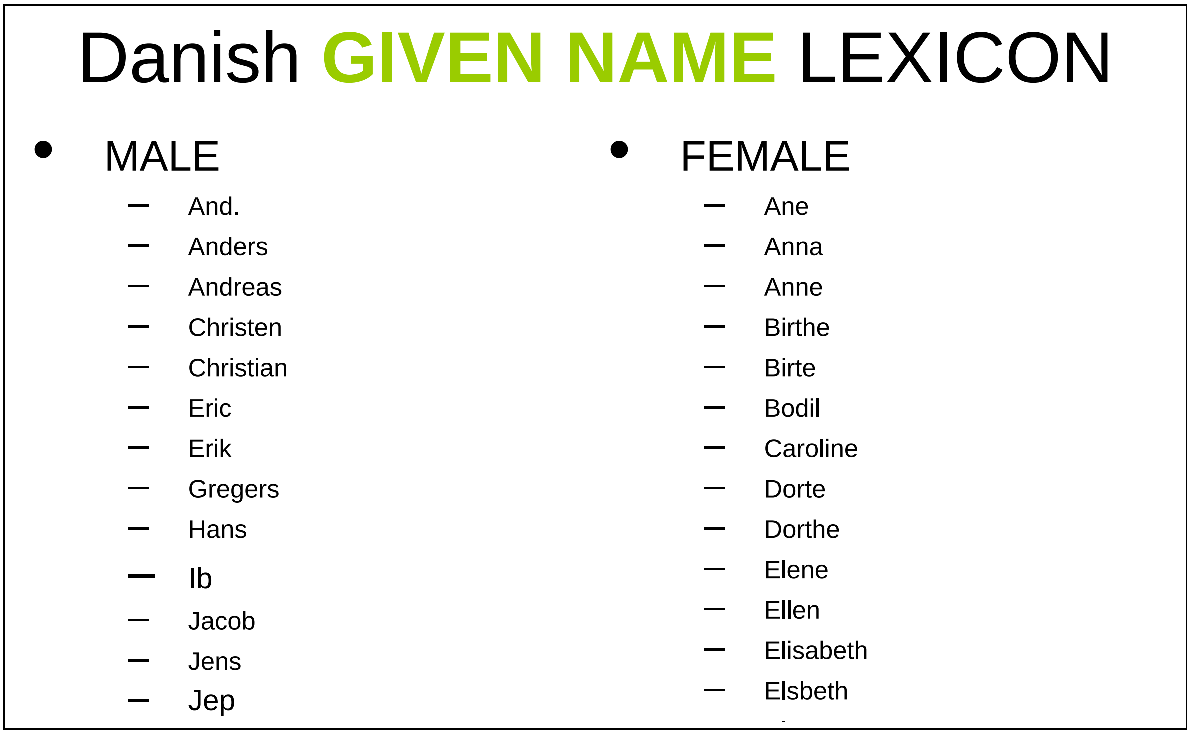
An ontology requires a precise description of what will be identified and semantically marked-up. As a beginning, these high-level entities have been chosen for the ontology:

• NAME	<NAME>
• DATE	<DATE>
• PLACE	<PLACE>
• RELATIONSHIP	<RELATION>
• OCCUPATION	<OCCUPATION>
• RECORD_TYPE	<RTYPE>
• SOURCE	<SOURCE>

These are basic semantic types that help to uniquely identify an individual in the record extracted. Eventually a standard will need to be developed for genealogical ontologies and their entities if wide-scale searching is going to be done across the whole semantic web. It may be that many ontologies will be developed for family history, but each will be mapped to the same standard family history ontology.

These descriptions need to be precise and understandable by the computer. For example, the 'RECORD_TYPE' is generally identified by its position in the record as the file name or the most immediate previous header – 'Christenings' or 'Burials' or 'Probates'.

In some instances it is impossible to adequately describe an entity. In such a case, a lexicon can be used to list all possible examples. Figure 1 shows a partial lexicon for common Danish given names before 1900. This lexicon is



Danish **GIVEN NAME** LEXICON

- MALE
  - And.
  - Anders
  - Andreas
  - Christen
  - Christian
  - Eric
  - Erik
  - Gregers
  - Hans
  - Ib
  - Jacob
  - Jens
  - Jep
- FEMALE
  - Ane
  - Anna
  - Anne
  - Birthe
  - Birte
  - Bodil
  - Caroline
  - Dorte
  - Dorthe
  - Elene
  - Ellen
  - Elisabeth
  - Elsbeth

Figure 1. Partial lexicon for common Danish given names before 1900.

attached to the NAME entity in the ontology. These lexicons can be as long or as short as is needed.

3. Solutions for special genealogical problems

Finally expert logic and reasoning need to be added inside the ontology to handle:

- Conversion functions
  - Compute birthdate from age at the time of death
  - Compute day, month, year from feast dates
- Matching different name forms as the same person
- Matching place names to appropriate records for search

Semantic Contributions

How will research be improved with the addition of semantics? First, the computer can work in a smarter way with the semantic definitions added by understanding the context of the word in the genealogical record. Secondly the speed of the search works amazingly fast over a very large web page. Thirdly, the same ontology can be used to annotate any designated web page. Fourthly, the accuracy is very good, but only if the target search is adequately specific - especially specific . The more detailed the target is; the more precise the results are.

The time-consuming steps are:

- The building of the ontology, but this is a one-time cost that is easily justified and precedes any search.
- The building of the PLACE lexicon since it must include every town and farm name, but this again only needs to be done once for any geographical area.

Research Contributions

The contributions of this work are:

- The first genealogical prototype for the semantic web
- A practical demonstration of the superiority of the semantic web for future research
- A portal for family history research that could be easily expanded with:
  - Maps
  - Look-ups
  - Helps
  - Training
  - Other countries and states