Family History Technology Conference Workshop

Genealogy, Microdata, and Search Engines

Robert Gardner, Ph.D., Google, Inc.
Tony Ruscoe, Google, Inc.
Dave Barney, Google, Inc.
January 25, 2012

Google's mission statement reads:

> *Google's mission is to organize the world's information and make it universally accessible and useful.*

This includes genealogical information. Over the last year, we have been anxiously engaged in an effort to improve the organization and usefulness of genealogical data available on the web. Google's tools are already widely used by genealogists, from Google Search, Image Search, and Maps used to find information about ancestors, to Google Docs and Picasa used to help organize collected information, to YouTube and Google+ with hangouts for collaboration and sharing. Without specifically targeting genealogy, Google's products have become powerful tools in the industry. Yet, we can do more to improve Google's coverage of genealogical data.

Schema.org

One method Google and others in the search industry have used to organize the world's information is to propose web standards in the form of schemas for how information can be marked-up and shared on the web. Schema.org, a widely accepted industry standard supported by Google, Bing, Yahoo, and others, is a collection of such schemas. By using these schemas with the microdata format, websites' HTML can explains to search engines **what** the data is, and no longer just how to display it.

From the schema.org webiste:

> *Your web pages have an underlying meaning that people understand when they read the web pages. But search engines have a limited understanding of what is being discussed on those pages. By adding additional tags to the HTML of your web pages—tags that say, "Hey search engine, this information describes this specific movie, or place, or person, or video"—you can help search engines and other applications better understand your content and display it in a useful, relevant way. Microdata is a set of tags, introduced with HTML5, that allows you to do this.*

Schema.org is widely used. The White House recently announced a new job search tool for youth, building upon the schema.org JobPosting standard. Search on Google News for "schema.org" for more examples of industry adoption.

Historical-Data.org

Schema.org does provide standards that can be employed by genealogy content providers to describe their data (e.g. there is a "Person" schema defined), but they are insufficient in the context of genealogy. As a result, we have recently launched a schema.org extenstion specific to genealogical data at www.historical-data.org to provide all of these benefits to the genealogy industry. By proposing this standard for content providers, Google, as well as other search engines, can better find, index, and provide relevant search results to users.

This standard has already been adopted by

FamilySearch.org, Geni.com, and WeRelate.org. As a result, Google now can better understand the genealogical data provided by these sites and make better use of it when users are searching for it.

The historica-data.org schemas include the following entities:

- HistoricalRecord
- HistoricalEvent
- HistoricalFamily
- HistoricalPerson

Two of these schemas are direct extensions of existing schema.org schemas:

- HistoricalPerson extends schema.org Person
- HistoricalEvent extends schema.org Event

The "Historical" prefix on each of these schemas is intended to differentiate it from schema.org schemas already in use (e.g. Person, Event, etc.).

Let's look in more detail at one of these schemas, HistoricalPerson. The HistoricalPerson extends the schema.org Person object. The Person object already includes, among others, the following fields:

- name
- birthdate
- gender
- parents
- children
- siblings
- spouse
- nationality

It does not include:

- birth information (e.g. location) other than birthdate
- death information
- christening information
- marriage information

In addition, the information for parents, children, and spouse simply point to other "Person" object with these same limitations.

The HistoricalPerson object adds, among others, the following fields:

- burial
- death
- contributor
- sources
- events (to capture life events not specifically covered in the schema)
- modifiedDate

Furthermore, HistoricalPerson also overrides the following fields, changing their type from "Person" to "HistoricalPerson":

- parents
- children
- siblings

Replacement for Gedcom?

Some have asked how the historical-data.org schemas relate to Gedcom and if this is a proposed replacement or extension of the Gedcom standard. Furthermore, why would content providers use these schemas if they already use genealogical standards (e.g. Gedcom) as their data model?

Historical-data.org is not a replacement for Gedcom, but rather a standard for content providers to communicate with search engines. While the schemas are rich in describing genealogical data, this schema is not as comprehensive as Gedcom. In addition, Gedcom is not a web standard. Gedcom data may not fully translate into historical-data.org schemas. However, it is the intent that all historical-data.org data could be converted to Gedcom or other standards. Therefore, it is safe to assume that historical-data.org is a subset of the Gedcom standard.

To restate its purpose, historical-data.org is a collection of schemas proposed as industry standards for marking up genealogical data on the web for the primary purpose of search engines better understanding and organizing the data.

Just for Search Engines?

While the primary purpose for historical-data.org is for content providers to better explain their genealogical data to search engines, we anticipate a wide variety of additional applications within the industry that can enhance the genealogy experience for users.

To demonstrate this, we have developed a Chrome browser extension. For users with this extension installed, when browsing a webpage with historical-data microdata, the extension detects and parses the data, notifies the user of its presence, and provides a number of convenient applications like performing a web search on that data. Other possible uses for the extension would be perhaps exporting it to a common format like Gedcom or adding the data to a collection of genealogical notes used during research. This could just as easily be done as a Firefox extension or other browser plug-ins.

As adoption of the historical-data.org standard increases across the industry, we anticpate seeing other new applications developed in industry to leverage and use this data.

A New Standard

We propose the historical-data.org schemas as a new web standard for sharing genealogical data on the web. We call on content providers to implement these schemas on their websites for describing their genealogical data as FamilySearch.org, Geni.com, and WeRelate.org have done. We also call on all search engines, both industry specific vertical search engines, as well as general web search engines, to adopt this standard. We challenge developers and the genealogy industry as a whole to develop new applications and uses of genealogical data described on the web in a common standard. We believe this will move the industry forward and ultimately provide better, more organized data to users on the web.