HTML5 Microdata Markup for Genealogy Sites

Family History Technology Workshop @rootstech 2013
Robert Gardner, Ph.D. and Dave Barney, Google, Inc.

HTML5 Microdata

There are currently several billion web pages containing genealogical information about historical people, places, and events. Most of it is unstructured, making it difficult to access by search engines and third-party tools. Search engines have made great strides in extracting useful content from unstructured web pages to enable searching, but extracting content in a way that provides meaning and structure has been elusive. Many efforts, such as The Semantic Web (http://semanticweb.org), have pioneered efforts to provide machine-readable structure to the unstructured content of web pages.

Early efforts at providing structure included the RDF standard used by The Semantic Web, which allows semantic linking between objects. RDFa was developed to allow RDF information to be embedded in HTML pages in the form of tag attributes. Later, microformats (http://microformats.org) was invented to provide not just linkage relationships but a rich vocabulary for describing various common objects such as people and events. When the HTML5 specification was introduced, it included yet another mechanism for adding semantic structure to HTML, named microdata (http://www.w3.org/TR/microdata/). Microdata is intended to be a simpler, more extensible mechanism for providing structured markup than previous attempts. Since it has the support of W3C, WHATWG, and the major browser vendors, it is expected that this format will eventually dominate the others. It is straightforward to automate the production of microdata in an existing web site and a variety of parsers are available that produce JSON from microdata. HTML5 includes a standard DOM API for retrieving and manipulating microdata, but it is not yet supported by all the major browsers.

Schema.org Vocabulary

While microdata markup provides a mechanism for adding structure to the data contained in a web page, it provides only the syntax or language structure. Semantic meaning is specified by various vocabularies built on top of the microdata. In the summer of 2011, Microsoft, Google, and Yahoo! formed a collaboration to create a standard set of vocabularies for many common objects. The collaboration, Schema.org, hosts many such vocabularies and accepts proposals for additions or enhancements to the vocabularies. In addition, the Schema.org vocabulary can be easily extended by cooperating organizations outside Schema.org.

Schema.org defines a hierarchical set of item types. The base type is `Thing` and all other types inherit the properties of `Thing`. A Schema.org `Thing` has four properties, `name`, `description`, `url`, and `image`. Additional specialized types specified by Schema.org include, for example, `Person`, `Place`, and `Event`. Schema.org defines types that are expected to have broad usage. Specific industries can use the extension mechanism to define types and properties for their needs.

Google, Bing, Yandex, and Yahoo! all recognize Schema.org types and enhance their search results based on the structure learned from this markup. In addition, markup can be used by a search engine to provide specific features. For example, Google Recipe Search uses markup to allow a user to filter search results by ingredients, cook time, and calories.

Historical-Data.org Extensions

In the spring of 2012, shortly after Schema.org was announced, several Google engineers working to improve Google tools for use in genealogical research formed a collaboration with FamilySearch and Geni to propose an extension to Schema.org for use in the genealogy industry. The resulting extension, historical-data.org, was published in time for RootsTech 2012. This proposal introduced extensions to standard Schema.org types that included additional historical information. For example, the extension included a type, http://historical-data.org/HistoricalPerson, which inherited from http://schema.org/Person and included a richer description of birth, death and marriage events.

The purpose of historical-data.org is to provide structure to genealogy web pages that search engines can use to enhance their results when users search for genealogical content and that tools can use to simplify genealogical research and cross-site collaboration. It is not intended as a complete data exchange language, so some items are missing that might be found in GEDCOM or other interchange standards.

In mid-2012, the historical-data.org extension was submitted to Schema.org for inclusion into the Schema.org vocabulary. The Schema.org working committee accepted the proposal but asked that it be folded directly into the Schema.org vocabulary. That is, instead of HistoricalPerson that extends Person, the new genealogy properties will be added directly to Person. Two new types, Family and HistoricalRecord, will be adopted as new Schema.org types. Since Schema.org types are extensible and implementations ignore types and properties they don’t understand, tools that use Schema.org can make use of the new types and properties if they choose, while existing tools will continue to work as originally designed.

The enhancements to Schema.org are scheduled for inclusion into the spec in the spring of 2013. Note, however, that due to the extensible nature of the Schema.org vocabulary, web sites that wish to publish genealogical information using the historical-data.org extensions can do so immediately, following instructions on http://historical-data.org, without waiting for the spec to be published by Schema.org. The full specification of additions and enhancements being added to Schema.org are documented on http://historical-data.org.

FamilySearch, Ancestry, Geni.com, MyHeritage, We Relate, and other companies have now adopted the historical-data.org microdata and are hosting web pages with the markup. We estimate that over 2 billion names are now available online with historical-data markup.

Reference Implementation

The authors have written a Chrome Browser Extension, People Finder, that recognizes historical-data markup on a web page, extracts the people described in the markup, and provides structured search into various search engines and genealogy sites. The extension provides simplified searches to Google's web search, image search, book search, and news archive search, structuring the queries in a way that maximizes the likelihood of retrieving relevant genealogy information. It also provides structured search for FamilySearch, We Relate, My Heritage, and Geni. The source code for this extension is in the process of being released into the open source community to serve as an example of how software can make use of the extensive number of names available with historical-data.org markup.

HTML Microdata identifies Items on a web page by adding attributes to the existing body HTML tags that indicate where Items start and end and what properties should be associated with the Item. An Item is indicated by adding an `itemscope` attribute to a tag. The Item's properties are found by gathering all tags in the enclosing scope that have an `itemprop` attribute. The `itemprop` attribute identifies a property of the Item. Each Item may also have a type, identified by the `itemtype` attribute that should accompany the `itemscope` attribute. Its value indicates the Item type as a URL that points to the place where the type is described. For example, the following HTML defines an Item of type "http://schema.org/Person" with properties `name` and `birthDate`:

```
<div itemscope itemtype="http://schema.org/Person">
  <span class="strong quiet" itemprop="name">
    Edward Montagu, 1st Earl of Sandwich
  </span>
  was born on <time itemprop="birthDate" datetime="1625-07-27">
    July 27, 1625</time>.
</div>
```

This HTML is written to produce the desired display for this person and the microdata is added to the HTML tags to define the structure. It is generally straightforward to add microdata with the same code that generates the HTML itself.

Now that we have an Item, we need a vocabulary to understand it. By visiting Schema.org we learn that an Item of type "http://schema.org/Person" represents a person, "alive, dead, undead, or fictional." There is a long list of properties associated with a person. The `name` property provides the person's name and the `birthDate` property provides the person's date of birth.

The People Finder Chrome Extension searches for Schema.org Person items on a page and, if it finds them, provides a popup window that displays the people and presents actions to search for them on various search engines. When one of those actions is clicked, a new tab is opened containing the search results. Currently, the extension provides structured search for FamilySearch, We Relate, My Heritage, and Geni. In addition, an action will search several of Google's corpora and produce results from web search, image search, book search, news archive search and (if birth or death locations are available) maps, structuring the queries in a way that maximizes the likelihood of retrieving relevant genealogy information.

The Chrome Extension is implemented in JavaScript. Although HTML5 defines a standard

JavaScript DOM API for accessing microdata, Chrome does not yet support this API, so the jQuery/microdata library is instead used to parse the microdata items from the HTML. To simplify communication between the various pieces of the Chrome Extension, a JSON representation of the microdata items is used. The JSON produced by jQuery/microdata for the sample HTML above is:

```
{
  "type": [ "http://schema.org/Person" ],
  "properties": {
    "name": [ "Edward Montagu, 1st Earl of Sandwich" ],
    "birthDate" : [ "1625-07-27" ],
  }
}
```

Note that the type and property values are all arrays since in microdata, these can all be multi-valued.

The Chrome Extension parses the JSON and places it into a specialized Person object to simplify building the user interface. The actions are simply link tags with a URL consisting of the proper host and query parameters to trigger the search. When an action is clicked, a new Chrome tab is opened at the link’s URL.

Conclusion

This paper has presented a collaboration among several of the major genealogy web publishers and several Google engineers to produce an HTML markup standard that can be used to improve search results and foster collaboration among web publishers.

Many of the largest web publishers are currently supporting the standard. FamilySearch is adding the markup to their entire collection. Geni already contains the markup. With their recent acquisition of Geni, My Heritage has committed to publishing the markup. Ancestry published the markup as part of the 1940 Census project. We Relate and others also include the markup in their published data. We estimate that as many as 2 billion pages currently use the historical-data.org markup.

For publishers, the primary benefit of using the markup at this point is to enhance search quality for genealogy-related searches. While no search engine currently makes explicit use of the markup to enhance the user interface, Google (and possibly others) provides enhanced search coverage for sites that contain the markup. Since the cost of producing the markup is minimal as it can be easily incorporated into existing templates and software, the authors encourage all genealogy web publishers to use the markup.

However, the true value of structured genealogy markup will be realized only as the developer community embraces it and begins writing tools that make use of it to produce tools for simplifying genealogy research and collaboration across genealogy sites. The authors have provided the People Finder Chrome Extension as an example of what can be done with the hope that it spur innovation and serve as a starting point for other developers.