MODULE 20

SHARING ARCHIVAL METADATA

Aaron Rubinstein

SOCIETY OF
American
Archivists

Appendix B: Case Studies

Implementing Schema.org at Rockefeller Archive Center

by Hillel Arnold, Rockefeller Archive Center

In October 2013, Rockefeller Archive Center (RAC) implemented Schema.org[47] markup in DIMES, its online discovery and access system. We did this largely in an effort to increase traffic to the site from web search engines. At that time, the site was acquiring less than ten percent of its total traffic from those sources, an abysmally low percentage.

In searching for ways to deal with this problem, I discovered that a number of other archives were experimenting with adding structured data to web pages describing their collections as a way of boosting their search relevance and rankings. Via the Code4Lib IRC,[48] I was able to connect with Will Sexton at Duke University and Scott Young at Montana State University, both of whom recommended that I add Schema.org metadata to finding aids. They also shared their expertise and documentation about implementing Schema.org at their institutions, which was a huge help and saved me countless hours and mistakes.

Schema.org is a collaborative effort between Google, Bing, Yandex and Yahoo! that aims to improve search results for all four companies. Because it is endorsed by the major search engines, it has quickly become the accepted way to embed structured data in web pages, which can then be interpreted and indexed by web crawlers. As others have noted, the properties available through Schema.org are more in line with the interests of businesses and other commercial entities than those of cultural heritage organizations, but there are ongoing attempts to add properties and types that fit the needs of libraries, archives, and museums.[49] Still, there are plenty of useful properties and types that can be used to help describe finding aids or other cultural heritage resources.

---

47 "Welcome to Schema.org," http://schema.org/docs/about.html, captured at https://perma .cc/3FGN-Y3ML

48 "IRC," Code4Lib, http://code4lib.org/irc, captured at https://perma.cc/4233-LQPX.

49 For example, the recent addition of a Libraries subtype (http://schema.org/Library, captured at https://perma.cc/FDK3-5JRE) and Schema Architypes (https://www.w3.org/community /architypes/, captured at https://perma.cc/U44S-9MX9).

Schema.org offers several ways to add structured data to a web page.[50] The approach RAC chose, because it was the easiest and quickest for us to implement, was to create a <div> with nested <meta> tags containing this data. This <div> does not display on a web page, but web crawlers can read and process that information, which leads to much more effective indexing.[51]

It is also possible to add Schema.org properties and types to existing HTML elements, but this requires you to structure your pages so that these properties are nested in accordance with how they are laid out in Schema.org.

Our discovery system, DIMES, is a heavily customized instance of the eXtensible Text Framework (XTF), an open source system developed by the California Digital Library that transforms XML documents to HTML using a series of XSL stylesheets. It was relatively easy to tweak the templates for finding aids and library records to include these tags, since these values are significant properties that were created according to DACS rules and encoded in EAD. In other words, consistency achieved through the use of content and structure standards enabled us to easily repurpose our existing archival description.[52]

For example, Figure 12 shows what the embedded Schema.org metadata looks like on the finding aid for the John D. Rockefeller papers.

When a web crawler sees these tags and attributes, it is able to apply semantic meaning to the data. It can find and correctly index important pieces of information about the page, such as its title, creator, and location. You can see what Googlebot sees on this page by dropping the URL into Google's Structured Data Testing Tool,[53] as the excerpt shows in Figure 13.

It took a few months,[54] but the improvement to our search rankings and relevance was significant. Traffic from search engines now accounts for more than a quarter of our total traffic, and many of our finding aids now appear on the first page of search results.

---

50 For detailed instructions, see the Lynda.com tutorial at http://www.lynda.com/Web -Content-Strategy-tutorials/Web-Semantics/143180-2.html, captured at https://perma .cc/8YDT-Y64W.

51 Although this example explains how to use Microdata to add Schema.org metadata to collection pages, embedded JSON-LD objects are emerging as the preferred format.

52 All code for our XTF implementation is available on Github at https://github.com /RockefellerArchiveCenter/XTF-RAC, captured at https://perma.cc/YZE9-JCFL.

53 Structured Data Testing Tool, Google.com, https://developers.google.com/structured-data /testing-tool/, captured at https://perma.cc/DC5R-T5DS.

54 To speed up the process, it helps to submit a site index, and or to take advantage of configuration tools for web crawlers available through Google, https://www.google.com /webmasters/tools/, or Bing, http://www.bing.com/toolbox/webmaster.

Figure 12. Schema.org Metadata

```
            <div itemscope="" itemtype="http://schema.org/CollectionPage">
                <meta itemprop="http://schema.org/description"
                    content="Material in this collection primarily documents John D. Rockefeller's
    activities during the 19th century. The collection is made up of correspondence, financial material,
    scrapbooks, and letterbooks. There are also several reels of audiovisual material at the beginning of
    the collection." />
                <meta itemprop="http://schema.org/name" content="John D. Rockefeller papers" />
                <div itemprop="http://schema.org/contentLocation" itemscope=""
                    itemtype="http://schema.org/Place">
                    <meta itemprop="http://schema.org/name" content="Rockefeller Archive Center" />
                    <meta itemprop="http://schema.org/url" content="http://rockarch.org" />
                    <div itemprop="http://schema.org/address" itemscope=""
                        itemtype="http://schema.org/PostalAddress">
                        <meta itemprop="streetAddress" content="15 Dayton Avenue" />
                        <meta itemprop="addressLocality" content="Sleepy Hollow" />
                        <meta itemprop="addressRegion" content="NY" />
                        <meta itemprop="postalCode" content="10591" />
                    </div>
                    <div itemprop="http://schema.org/geo" itemscope=""
                        itemtype="http://schema.org/GeoCoordinates">
                        <meta itemprop="http://schema.org/latitude" content="41.091845" />
                        <meta itemprop="http://schema.org/longitude" content="-73.835265" />
                    </div>
                    <meta itemprop="http://schema.org/telephone" content="(914) 366-6300" />
                </div>
                <meta itemprop="http://schema.org/dateCreated" content="1855-1942 (bulk 1879-1894)" />
                <meta itemprop="http://schema.org/inLanguage" content="en" />
                <div itemprop="http://schema.org/publisher" itemscope=""
                    itemtype="http://schema.org/organization">
                    <meta itemprop="http://schema.org/name" content="Rockefeller Archive Center" />
                    <meta itemprop="http://schema.org/url" content="http://rockarch.org" />
                    <div itemprop="http://schema.org/address" itemscope=""
                        itemtype="http://schema.org/PostalAddress">
                        <meta itemprop="streetAddress" content="15 Dayton Avenue" />
                        <meta itemprop="addressLocality" content="Sleepy Hollow" />
                        <meta itemprop="addressRegion" content="NY" />
                        <meta itemprop="postalCode" content="10591" />
                    </div>
                    <meta itemprop="http://schema.org/telephone" content="(914) 366-6300" />
                </div>
            </div>
```

Figure 13. Google's Structured Data Testing Tool

CollectionPage	0 ERRORS 0 WARNINGS ^
@type	CollectionPage
description	Material in this collection primarily documents John D. Rockefeller's activities during the 19th century. The collection is made up of correspondence, financial material, scrapbooks, and letterbooks. There are also several reels of audiovisual material at the beginning of the collection.
name	John D. Rockefeller papers
dateCreated	1855-1942 (bulk 1879-1894)
inLanguage	en
contentLocation	
@type	Place
name	Rockefeller Archive Center
url	http://rockarch.org
telephone	(914) 366-6300
address	
@type	PostalAddress
streetAddress	15 Dayton Avenue
addressLocality	Sleepy Hollow
addressRegion	NY