Gateway to Oklahoma History Case Study:
Structured Data and Metadata Evaluation for Improved Image Resource Findability on the Web

Emily Ann Kolvitz
University of Oklahoma,
USA
kolvitz1@gmail.com

Keywords: structured data; metadata; Semantic Web; online search; information retrieval; image findability

1. Introduction

Image Resource Findability on the World Wide Web is still very much a land-grab. For the Semantic Web to become a reality online businesses and individuals have to get their hands dirty and also come face-to-face with the realization that search engine giants are increasingly becoming the go-to tool for information resource retrieval. “Increasingly, students use Web search engines such as Google to locate information resources rather than seek out library online catalogs or databases of scholarly journal articles” (Lippincott 2013). This puts the search engine giant in a unique position to dictate how the future of search will work on the Web - and therefore, your organization’s future presence (or lack thereof) on the Web. Search Engine Optimization (SEO) techniques change frequently and remain much a mystery to many companies. The one variable in the equation of Web findability that remains a staple is good quality metadata under the hood of the Website. In this case study, a methodology is applied to the Gateway to Oklahoma History’s Website. This study can be generalized to organizations looking to benchmark their own findability maturity on the Web from an image-centric viewpoint.

2. Purpose

Image search and retrieval is a more difficult area than text search and retrieval because accessibility to the image content is largely dependent on the context presented in and around the image resource. The future of Semantic Web technologies relies very much on the idea that organizations are fluent in structured data and have devoted resources to exposing valuable data to the web. The W3C (World Wide Web Consortium) was founded over two decades ago and the widespread adoption of Schema.org and structured data on the Web did not gain traction until the big four search engines (Google, Bing, Yahoo, and Yandex) agreed that a standard was needed to pave the way forward. “On-page markup helps search engines understand the information on web pages and provide richer search results. A shared markup vocabulary makes easier for webmasters to decide on a markup schema and get the maximum benefit for their efforts.” (Schema.org 2014) Many organizations still lag behind on their implementation of any type of structured data. Structured data is only one piece of the findability algorithm. Metadata near content, embedded within content, or listed in the alt text of an html document all tell machines something about the content inside of the record as well. There are no guardians of the Web, ensuring structured data is uniformly applied to all records with equal attention and care and there is no standard, mandated requirement for records on the Web to provide context for image resource findability. Most search engines do not crawl embedded XMP data or the invisible Web, leaving text near images, file names or text in the alt-text in html markup as the only context for image resources. The search algorithms for image retrieval are subject to change frequently (Kritzinger 2013) and additionally, social media sites and organizations strip embedded data from images (Embedded Metadata Manifesto 2014). Embedded metadata provides context and
provenance for image resources. Even with the dramatic adoption of structured data markup utilizing schema.org vocabularies, there still remains metadata opportunities on the Web. Reicks recommends embedded metadata as a strategy for online findability by showcasing examples of applications that parse embedded data into structured data around images on the Web such as PhotoShelter and LicenseStream (2010).

2. Research Methods

The following research question informed this project: What are the types and quality of structured data, XMP, and metadata records available for image resources appearing on the website? Utilizing the Structured Data Linter Tool and Phil Harvey’s ExifTool, information was gathered to quantify these research questions. Image records on the Gateway to Oklahoma History’s website were investigated for the types, quality and quantity of embedded metadata and structured data.

3. Results

The Gateway to Oklahoma History’s Website has a wealth of structured data and metadata pertaining to its image resources. Search queries utilizing structured data markup tags and/or embedded metadata yielded relevant and accurate results during a normal web search, but did not yield relevant and/or accurate image resources during an image search. Descriptive filenames were not used for image resources, which is an important part of image retrieval through web search engines. Adding Schema.org tags to the on-page markup, to accompany the structured data already present is another area for improvement. An interesting finding from this research was that embedded metadata was only found on the largest, original version of the image resource, and never on smaller derivative images. Structured data included in the on-page markup included Open Graph Protocol and Dublin Core. IPTC was the primarily type of embedded metadata present for the image resources.

4. Conclusion

The results and methodology for this research can help GLAM institutions (Galleries, Libraries, Archives & Museums) by bringing awareness to the state of structured data and image resource findability for cultural heritage institutions on the Web. GLAMs must be active in the SEO space, support machine-readable language in the markup of their sites, and utilize Schema.org vocabularies and descriptive filenames for relevancy in search engine results. The Digital Library Federation, which is a program of the Council on Library and Information Resources, concludes that “Getting found means repository objects must be included in the indexes of major search engines because most students and faculty now begin their research with Internet search engines. Digital repositories created by libraries will be largely invisible to users if their contents are not indexed in these search engines” (Digital Library Foundation 2014).

References


