The Web Observatory Extension: Facilitating Web Science Collaboration through Semantic Markup

Dominic DiFranzo
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
difrad@rpi.edu

John S. Erickson
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
erickj4@rpi.edu

Marie Joan Kristine T. Gloria
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
glorim@rpi.edu

Joanne S. Luciano
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
jluciano@rpi.edu

Deborah L. McGuinness
Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
dlm@cs.rpi.edu

James Hendler
Tetherles World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA
handler@rpi.edu

ABSTRACT
The multi-disciplinary nature of Web Science and the large size and diversity of data collected and studied by its practitioners has inspired a new type of Web resource known as the Web Observatory. Web observatories are platforms that enable researchers to collect, analyze and share data about the Web and to share tools for Web research. At the Boston Web Observatory Workshop 2013 [3], a semantic model for describing Web Observatories was drafted and an extension to the schema.org microdata vocabulary collection was proposed. This paper details our implementation of the proposed extension, and how we have applied it to the Web Observatory Portal created by the Tetherless World Constellation at Rensselaer Polytechnic Institute (TWC RPI). We recognize this effort to be the "first-step" in the construction, evaluation and validation of the Web observatory model and not the final recommendation. Our hope is that this extension recommendation and our initial implementation sparks additional discussion among the Web Science community of on whether such direction enables Web Observatory curators to better expose and explain their individual Web Observatories to others, thereby enabling better collaboration between researchers across the Web Science community

Categories and Subject Descriptors
K.4.0 [**Computers and Society**]: General

Keywords
Web Observatory, Web Science.

1. INTRODUCTION
As the World Wide Web continues to evolve and greatly impact our everyday lives, researchers from a wide array of perspectives and disciplines have turned their attention to its study. However, while the research, analysis and data are plentiful, those who study the Web lack a more united understanding of the Web's influence. The field of Web Science was developed in response to this, to help create a coherent and connected study of the Web [1].

WWW'14 Companion, April 7-11, 2014 Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576936

Still in its infancy, the field of Web Science now faces several unique challenges from conflicting methodological philosophies to developing a suitable infrastructure for its scientists to collaborate and share resources. More importantly, the scale and scope of the data produced by users of the Web and studied by Web Scientists is unprecedented for any modern field of study. Web Scientists are thus looking to mixed methodology practices, new computational infrastructure and large scale analytics in order to better make sense of this complex phenomena.

As a means for managing this research complexity, the Web Science community has undertaken the development of a new distributed platform to facilitate the collection, analysis and sharing of data about the Web. Termed a "Web observatory," the accepted definition of this platform is:

> "a distributed archive of data on the Web and its activity, and, at the same time, mechanisms and tools that will be able to explore its development in the past, to examine its present condition and to establish potential developments in the future" [4].

Based on this definition, Web observatories promise to be a vast improvement over current resources used by Web scientists, which are centralized, hard-to-find and often proprietary

The Web Science Trust's Web Observatory Project[1] is motivated by three goals: First, to create a global data resource that moves beyond the traditional understanding of a centralized data warehouse to that of a more distributed environment for interdisciplinary analysis and knowledge sharing. Second, to provide Web scientists a space to foster the development and sharing of toolsets, frameworks and workflows. This can only be accomplished by adopting a bottom-up approach that aggregates individual repositories into a virtual infrastructure. Finally, the Web Observatory Project aims to promote and empower researchers to use not just quantitative correlation methods on datasets, but to explore and incorporate qualitative analyses that

[1] For additional information the Web Science Trust Web Observatory Project, please see http://webscience.org/web-observatory/

may help provide a more comprehensive understanding of the socio-technical evolution of the Web [5].

The first milestone in the realization of the Web Observatory Project was the development of the Web Observatory Wiki[2], based on the Semantic MediaWiki[3] platform and created by Jérôme Kunegis at the Institute for Web Science and Technologies of the University of Koblenz–Landau. The Web Observatory Wiki is a curated list of the known Web observatories, datasets, and organizations engaged in the larger Web Observatory effort. The Semantic MediaWiki infrastructure implements a concise schema to help categorize and organize the metadata attributes around datasets, dataset repositories, and organizations.

The Web Observatory Wiki is an important first step towards aggregating information about current Web observatories and facilitating their discovery; however, there are many areas of improvement that need to be considered. Critically, as a Semantic MediaWiki instance, the Web Observatory Wiki remains a centralized repository of Web observatory metadata and information. Moreover, the Web Observatory Wiki also lacks fields and attributes for entities that are not just datasets. The Web Observatory Project on the other hand features tools and methods in a decentralized environment that the Web Science community can contribute to, interlink to, view and explore.

On October 9, 2013, a large group of Web Science researchers from universities in the Web Science Trust came together for the Boston Web Observatory Workshop (WOW) [3]. The goal of Boston WOW was to identify and discuss further issues in Web observatory development and, more importantly, to agree on next steps. An important result of the workshop was an agreement by contributors to collaborate on defining a schema.org vocabulary extension for Web Observatories [6]. By enabling providers to embed metadata in Web resources in ways that are more easily indexed by major search engines and other services, a Web Observatory schema.org extension presents a more lightweight, flexible way to create connections between Web observatories without the need for constructing a formal structure this early in the project's development The remainder of this paper presents the work of our team in defining the Web Observatory extension and in implementing and testing it within our own RPI WSRC Web Observatory portal.

2. SCHEMA.ORG AND MICRODATA

schema.org is an initiative launched by the leading search engine providers to create and support a common set of schemas for structured data markup on Web pages using a particular format known as *microdata*.[4] These standardized vocabularies enable the metadata to be more machine readable, allowing for software agents to better search, discover and display this information. The set of schema.org extensions recognized by the community has grown steadily since 2011 and now includes a wide range of topic categories and resource types [7]. Most notably this includes the Dataset extension, useful for describing datasets and data catalogs published on the Web.[5]

According to a study on the 2012 Web corpus published by the Common Crawl foundation, 6.1% of all websites that have structured data use microdata [2]. Of those sites that use microdata and are in the Alexa Top 1000 list (meaning they are the top websites on the Web), 31.67% use microdata [2]. Some of these top websites that use microdata are Apple, Microsoft, and eBay.

2.1 Goals of Web Observatory Schema.org Vocabulary

The Boston WOW group recognized that in order to realize its objectives, a schema.org extension recommendation would be a necessary next step. The Web Science Research Center at RPI and our collaborators have proposed extending the schema.org vocabulary to include attributes and properties that can be used to describe a variety of Web observatories. The community chose to extend schema.org because it provides sufficient structure through standardization without the need of formal definitions at this point. To reiterate, the Web Observatory Project is in its infancy; and, the consensus to explore a schema.org vocabulary extension is an attempt by the community to realize and definite Web observatories while presenting a reflexive framework for evaluation. The following goals influenced our team's definition of this vocabulary.

2.1.1 Describe Web Observatories

First and foremost, the Web Science community needs a more formalized mechanism for defining, organizing, and expressing metadata to be used for describing Web observatories. Because the concept of a Web observatory is still in its infancy, we saw this work as an opportunity to gather consensus on a more formal, structured, standard definition for *what constitutes a Web observatory* and *how to communicate* this information to others, including humans and software agents. The resulting standard would need to be sufficiently open-ended to include the wide variety of datasets, tools and methods used in Web Science, but without conceding meaning or purpose.

2.1.2 Interconnect Web Observatories

Many current Web observatories are scattered and siloed from each other. Therefore, in addition to describing Web observatories in a standard way, the community must start interlinking these observatories more effectively. Considering the scale of data involved, this seems to be a daunting task. However, we consider the schema.org Web Observatory vocabulary extension to be a practical first step in helping to locate and highlight potential links between Web observatories that implement the standard. By making Web observatory metadata more explicit and more easily indexed by search engines and other agents, the Web Science community can better identify opportunities for interlinking related Web observatories.

2.1.3 Facilitate discovery of tools, datasets, and projects for researchers

A primary goal of the schema.org Web Observatory extension is to enable researchers to more easily discover and acquire access to tools and datasets from other Web Science researchers, thereby

[2] See, http://wow.west.webobservatory.org/index.php/Main_Page for additional information on the Web Observatory Wiki

[3] Details on the Semantic MediaWiki can be found at http://semantic-mediawiki.org/

[4] For the full definition, please visit http://schema.org

[5] For a full listing of the schema.org dataset, see http://schema.org/Dataset

accelerating collaboration and innovation. By exposing the metadata used to describe Web observatories and their contents in a standard and machine readable way, Web Observatories become less opaque and better tools can be built to support the research tasks of using data, understanding it and presenting it to users.

3. SCHEMA.ORG VOCABULARY

The initiate draft of the schema.org Web Observatory extension proposes four new classes: *Web Observatory*[6], *Web Observatory Project*[7], *Web Observatory Tool*[8] and *Web Observatory Dataset*[9]. An advantage of the schema.org extension model is that many of the properties we need to describe these new classes (keywords, description, author, etc.) are inherited from pre-existing base classes. Our team was able to focus on only those properties that were unique to Web observatories. In the following sections, we describe in further detail the new added classes as well as include a simple example.

Thing > CreativeWork > Web Observatory

This is a base class describing a Web Observatory. It is a subclass of `CreativeWork`, which includes many of the properties we already needed (such as name, keyword, description, author, etc). The additional property added is `webObservatoryProject`, which is useful for listing Web observatory projects that are included in a particular Web observatory. The following is a simple example of the Web Observatory class in use.

```
<div itemscope
itemtype="http://schema.org/WebObservatory">
<span itemprop="name"><b>TWC Web
Observatory</b></span>
<meta itemprop="url"
content="http://tw.rpi.edu/web/TWCObs">

<span itemdrop="webObservatoryProjects">
      <span itemscope
itemtype="http://schema.org/WebObservatoryProjec
t">
         <span itemprop="name">
             <a
href="http://tw.rpi.edu/web/project/FirstRespond
ers"> First Responders Network Observatory</a>
          </span>
          <meta itemprop="url"
content="http://tw.rpi.edu/web/project/FirstResp
onders"/>
          </span>
     </span>
</div>
```

The full Web Observatory class definition proposal may be found via the TWC RPI Web Schemas project site.

Thing > CreativeWork > Web Observatory Project

This class is used to annotate Web observatory projects. Projects are defined by a name and include a collection of datasets, tools and methods. The added properties for this class are `webObservatory` (points to a Web observatory that a particular project belongs to), `webObservatoryTool` (a tool that is used in this project), `webObservatoryDataset` (a dataset used in this project) and `method` (a method used in this project, i.e. Social Network Analysis). The following is a simple example of the Web Observatory Project class in use.

```
<div itemscope
itemtype="http://schema.org/WebObservatoryProjec
t">

     <span itemprop="name">

       <b>TWC International Open Government
Dataset Search (IOGDS)</b>

      </span>

      <meta itemprop="url"
content="http://logd.tw.rpi.edu/page/internation
al_dataset_catalog_search"/>
```

The full Web Observatory Project class definition proposal may be found via the TWC RPI Web Schemas project site.

Thing > Creative Work > Web Observatory Dataset

This class is a subset of the schema.org `Dataset` class. It includes all the same properties of schema.org/Dataset, but also includes `webObservatoryProject`, which points back to the a specified Web observatory project. This class also includes `accessPolicy`, which states whether this dataset has open or closed access. The following is a simple example of the Web Observatory Dataset class in use:

```
<div itemscope
itemtype="http://schema.org/Dataset/WebObservatory
Dataset">
        <span itemprop="name">
           <b>International Open Government
Dataset Search Metadata</b>
         </span>

            <meta itemprop="url"
content="http://purl.org/twc/vocab/conversion/Me
taDataset"/>
</span>
</div>
```

The full Web Observatory Dataset class definition proposal may be found via the TWC RPI Web Schemas project site.

Thing > CreativeWork > Web Observatory Tool

This is a class that is used to annotate tools developed for and used by Web observatories. This may include visualizations, data converters and more. The following is a simple example of the Web Observatory Tool class in use.

```
<div itemscope
itemtype="http://schema.org/WebObservatoryTool">

        <span itemprop="name">

            <b>S2S</b>

        </span>

        <meta itemprop="url"
content="http://tw.rpi.edu/web/project/sesf/work
inggroups/s2s"/>

    </div>
```

[6] Full description for "Web Observatory" can be found at http://logd.tw.rpi.edu/web_observatory

[7] Full description for "Web Observatory Project" can be found at http://logd.tw.rpi.edu/web_observatory_project

[8] Full description for "Web Observatory Tool" can be found at http://logd.tw.rpi.edu/web_observatory_tool

[9] Full description for "Web Observatory Dataset" can be found at http://logd.tw.rpi.edu/web_observatory_dataset

The full Web Observatory Tool class definition proposal may be found via the TWC RPI Web Schemas project site.

4. SCHEMA.ORG IMPLEMENTATION

The Tetherless World Constellation's Web Observatory Portal [http://tw.rpi.edu/web/web_observatory] (see Figure 1) provides a reference implementation of all levels of the Web Observatory class hierarchy. This portal was the basis of our evaluation using the Google Structure Data Testing Tool[10] and the Yandex Structured Data Validator.[11]

5. TESTING AND EVALUATION

The proposed schema.org Web Observatory extension will be introduced to the public via the W3C Web Schemas Community and Web Observatory Community in December 2013. Examples of all classes of the vocabulary have been instantiated on the TWC RPI Web Observatory Portal and were tested against the Google Structured Data Testing Tool and the Yandex Structured Data Validator. Both services successfully recognized and parsed both the schema.org microdata and RDFa 1.1 Lite version of the extension, demonstrating that the terms are correctly instantiated on our portal.

This demonstrates that our extension conforms to the technical specifications of schema.org microdata and RDFa 1.1 Lite. The next step is to t also evaluate whether our extension meets the goals we outlined above. Again, the three main goals of this Web observatory extension were: 1) to describe Web observatories, 2) interconnect Web observatories together, and 3) to facilitate discovery of tools, datasets, and projects for web science researchers. To evaluate whether the WOW 2013 model and our implementation of it currently meets these goals, we need to develop use cases and show that they are met. Currently, an evaluation can only be completed against the first goal, describing Web observatories. The Web observatories in the TWC Web Observatory portal are described completely using our implementation of the schema.org extensions. However, we recognize that further evaluation of goal one will be revisited as more Web observatories in the future try to use our implementation. Goals two and three are largely dependent on adoption and implementation of the schema.org extension by other labs. Therefore, we recognize these evaluations to be a reflexive and iterative process. We can only evaluate our methods and efficacy of Web observatories as they are built, used and shared by the community.

In addition to evaluating the goals we stated before, we believe the efforts of the WOW community must focus on several key issues:

1. **Does the web observatory information model proposed at WOW 2013 adequately cover a variety of Web observatory uses?** This again ties in strongly to our first goal in describing Web Observatories. We will address this by soliciting use cases covering diverse examples of Web observatories, implemented or planned, and mapping descriptions to the information model. We plan on this model growing and changing over time as we become exposed to the diversity of Web observatories our community creates.

[10] To evaluate on via Google, please use http://www.google.com/webmasters/tools/richsnippets

[11] To evaluate using Yandex, please use http://webmaster.yandex.com/microtest.xml

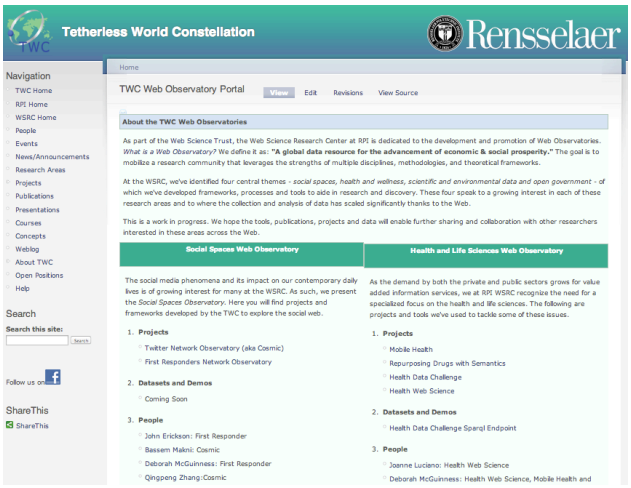

Figure 1: TWC Web Observatory Portal

2. **Does the schema.org Web Observatory extension correctly implement the WOW 2013 model?** The current paper documents one implementation of the proposed extension; its authors encourage discussion and feedback from the WOW community.

3. **What barriers to adoption are revealed as the community attempts practical implementation?** Testing of the Web Observatory extension on a diversity of platforms may indicate that certain elements are difficult to implement. Furthermore, practical implementation may suggest specialized tools (including code libraries) that will make implementation of the Web Observatory extension easier on some platforms.

At the time of this submission, we are confident that our draft implementation of a schema.org Web Observatory extension, and the Web Observatory information model it is based on, will address the goals of the Web Science community in describing and interlinking Web observatories and facilitating the discovery of tools, datasets, and projects for Web Science researchers.

6. FUTURE WORK

This paper has presented the initial draft of our proposed schema.org Web Observatory vocabulary extension. Through out the paper, we discussed what has been developed and implemented within our own lab in order to realize some of the goals of the overall Web Observatory Project. However, many gaps remain and there is a substantial need for community feedback in order to achieve the remainder of the project's goals. In this section, we present additional next steps to be considered as well as long-term projects that may help breach the gaps in goals two and three. As stated before, we need to evaluate whether our implementation currently captures the proposed WOW 2013 information model. This includes conducting in-depth discussions with the members of WOW 2013 and encouraing the implementation our vocabulary extenions within their own Web Observatory portals. Outside of the WOW 2013 group, we need to evaluate that the proposed model information model can cover the meta-data needs of the many Web observatories that also exist. We plan to engage the greater Web Observatory and Web Science communities through webinars and future workshops. The results of this outreach we hope will both expand and grow the extension along with increase it's use in other Web Observatories.

As more Web Science research groups adopt the proposed vocabulary, we look forward to realizing the other goals of this extension. With this extension in place and in use, new federated search applications, similar to TWC's International Open Government Data Search application[12], can start to be realized. We can also start to build Web agents that can query, crawl and better understand the nature of Web observatories as they are being used, giving us recommendations and ideas on how these Web observatories could be better linked and used.

We also look forward to exploring other extensions to schema.org to better address issues in Web data that are of concern to Web scientists. For example, it might be useful to extend the notion of Place (location) to include virtual places. In existing vocabularies, there is no way to address the concept of a virtual community or social network other than as a Web application or Web page. In Web Science, we treat these online communities as having a sense of "virtual" space from where our data flows and where the users we study interact and meet.

Based on our experience, we believe the community also needs better tools to assist in annotating Web Observatory pages with schema.org microdata. Our reference annotations using the Web Observatory extension were accomplished manually or with the help of special purpose code. A more generic and easy-to-use tool would enable more Web Science research labs to integrate and expose their Web observatory metadata using the schema.org Web Observatory extension.

7. CONCLUSION

To better enable the goals of the Web Observatory project, better exposure of metadata in a machine readable format is needed. This work is a draft of an extension to schema.org to enable this. We've outlined the goals we were seeking to reach, along with a clear explanation of what exactly our extension entails. With future work, help and adoption from the rest of the Web Observatory community, we hope to reach these goals.

8. ACKNOWLEDGEMENTS

Our thanks to the Tetherless World Constellation lab for their support and help in the development of tools featured in the TWC Web Observatory Portal. We also wish to give additional thanks to the entire Web Science Community for their recommendations and feedback on the schema.org extension vocabulary.

9. REFERENCES

[1] Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., & Weitzner, D. 2006. Creating A Science Of The Web. *Science* Vol. 313 no. 5788 (11 August 2006), 769-771. DOI:10.1126/science.1126902

[2] Bizer, C., Eckert, K., Meusel, R., . . . Volker, J. 2013. Deployment of RDFa, Microdata, and Microformats on the Web. In proceedings *International Semantic Web Conference.* Sydney, Australia. (2013 Nov). https://bitly.com/shorten/

[3] Brown. I. 2013. Boston Web Observatory (Web Observatory Workshop) October 9 2013. Website. (2 Oct. 2013). http://bit.ly/1hRTztf

[4] Gallen, C. 2012. A definition of the Web Observatory. Website. (24 July 2012). http://bit.ly/18Hce2F

[5] Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N., & Hendler, J. The Web Science Observatory. *IEEE Intelligent Systems*, vol. 28, no. 2. (March-April 2013) 100-104. DOI:10.1109/MIS.2013.5 SOTON: http://bit.ly/1crWAsP.

[6] Technical Discussion. Boston Web Observatory Workshop. 2013. Shared Google Doc. http://webscience.org/boston-wow-web-observatory-workshop-oct-9th-2013/

[7] W3C. (n.d.). Proposals for schema.org: Already accepted and added. Website. http://bit.ly/1f7pSjS

[12] Additional information on IOGDS, please visit http://logd.tw.rpi.edu/page/international_dataset_catalog_search