Building Semantically-Enriched Web Observatories

Marie Joan Kristine Gloria[1], Joanne S. Luciano[1], Deborah L. McGuinness[1]
[1]Web Science Research Center, Rensselaer Polytechnic Institute. 110 Eighth St.,
Troy, NY USA 12180
{glorim@rpi.edu, jluciano@rpi.edu, dlm@cs.rpi.edu}

Abstract. We describe selected work conducted by Rensselaer Polytechnic Institute's Web Science Research Center (RPI WSRC) aimed at building web observatories. Specifically, we discuss the research center's best practices for building semantic-enabled technologies and automated processes to discover, collect, analyze, and interpret large amounts of data. To illustrate this, we present two toolsets, built to address issues of data collection and management within our Health and Life Sciences Web Observatory and Social Spaces Web Observatory.

Keywords: Semantic technologies, Web Observatories, Web Science, Social Spaces, Health and Life Science

1 Introduction

One view of a "web observatory" is a repository that represents one or more aspects of the World Wide Web structured such that observations can be made, activity can be monitored, and experiments may be performed. To build and monitor these observatories, new tools and processes are needed that address the Web's complexity and multifaceted nature. This contribution advocates for the development and use of semantically-enriched tools that will ease generation of and exploration within Web observatories.

2 Automating Health Data

Health Web Science is an emerging subfield of Web Science which looks to understand the increasing amounts of data from the health and life science fields [1]. In order to advance Health Web Science, one approach is in automating health data. For example, the Health and Human Services (HSS) noted in 2011 that data are often inaccessible and incongruent, making it difficult to use. In an attempt to address these challenges, HHS sought help from the developer community by initiating challenges. RPI TWC responded and was awarded first place in the Metadata developer challenge providing tools that enabled the discovery of, access to, and integration of the HHS's 368 datasets [4].

To specifically address this challenge, we used the csv2rfd4lod-automation tool developed at RPI TWC to aggregate and integrate across multiple versions of multiple datasets of multiple source organizations in an incremental and backward-compatible way [3]. We mirrored the hub.healthdata.gov CKAN instance using its API to our own instance at healthdata.tw.rpi.edu/hub. This allowed us to both improve the CKAN-based metadata, including adding Data Dictionaries and Technical Documentation as Resources, and to improve the RDF generated by CKAN. The datasets were organized according to "SDV" -- their *Source* organization, *Dataset* identifier, and the dataset *Version.* We then used LODSPeaKr, another RPI TWC tool, designed to create Linked Data applications and publish RDF data quickly and with minimal effort in order to improve data accessibility for humans and for machines [3]. Coupled with the csv2rfd4lod-automation tool, the TWC developed a streamlined, replicable process to convert and enhance metadata of the HHS datasets. Our work resulted in over 12 million triples loaded from 1.9 billion triples creating 135 abstract datasets, 238 versioned datasets, and 129 layered datasets[1] [2].

3 Structuring Social Data

The social Web introduces yet another layer of study for Web Science. As such, there is an increasing need to develop tools and methods that can be used by a diverse range of researchers interested in understanding human social behaviors.

The Twitter Network Observatory project seeks to build a semantically-enabled platform that aggregates, stores, and analyzes Twitter data. The data are converted to RDF linked data and re-published via a TWC LOGD SPARQL endpoint. This enables researchers the ability to access the data and explore the relationships of people and semantics within the graph database. In addition, users can visualize and analyze different types of sub-graphs based on selections of topic, network definition, time range, sentiments, location, etc. The Twitter Network Observatory performs a series of quantitative analyses to explore the topological properties of the extracted social graphs. In addition, the network files can be exported for use by other toolkits. One of our funded applications[2] is using Twitter content to support requirements gathering for first responders using social media. The project is developing tools to identify appropriate *hashtags* and social media topic-appropriate contributors. The project is also analyzing tweet networks to support the requirements gathering process. While this effort focuses on

[1] For more information regarding granularities, please visit <https://github.com/timrdf/csv2rdf4lod-automation/wiki/Dataset-granularities:-Abstract-vs.-Versioned-vs.-Layer>

[2] Additional information on the First Responders Project: <http://tw.rpi.edu/web/project/FirstResponders>

first responders, the tools being developed are useful for gathering requirements and for analyzing emerging social media generated graphs.

It should also be noted that the Twitter Network Observatory was developed independently of the methods and processes used in the HHS Metadata Challenge. We recognize the strengths and weaknesses from each project and will continue to learn from these endeavors. We intend to incorporate and build-upon each initiative for future Web observatories.

4 Conclusion

We have briefly introduced two toolsets used to generate, analyze, and interact with web observatories and their content. The tools embody our continued commitment to design and build semantically-enriched tools to aide in the aggregation, integration, and analysis of large data sets. In turn, researchers are empowered to explore data and to engage in further collaborative efforts. However, we recognize and offer for further discussion, additional challenges in developing the correct tools and methods for spaces like the social Web. We believe that Web observatories serve as exemplary collections of tools, methods, and provide case studies for others to leverage; however, we are yet to fully consider the limitations of such observatories in light of legal, ethical, and social concerns.

Acknowledgments. Our thanks to all of the researchers, faculty, and staff of the Tetherless World Constellation at RPI. We would like to thank Jim McCusker, Tim Lebo and Alvaro Graves for their contribution to the HHS Data Challenge. In addition, we would like to thank Bassem Makni, Qingpeng Zhang, John Erickson, Rui Yan, Katie Chastain, and Zach Fry for their assistance with aspects of the Twitter Network Observatory.

5 References

1. Brooks, E. H., Cumming, G. P., & Luciano, J. S. Health web science: application of web science to the area of health education and health care. In Proceedings of the second international workshop on Web science and information exchange in the medical web. (pp. 11-14). ACM; 2011
2. McCusker, J., Lebo, T., Graves, A., and Gloria, M. (2013). “Health Data Statistics”, http://healthdata.tw.rpi.edu/statistics
3. McCusker, J., Lebo, T., Graves, A., and Gloria, M. (2013). “Wiki: TWC-HealthData”. Github, https://github.com/jimmccusker/twc-healthdata/wiki
4. Wong, A. (2013). “Winners of Health Data Platform Challenge.” Health 2.0 Website. 20 Feb. 2013, http://www.health2news.com/2013/02/20/winners-of-health-data-platform-challenges/