Exploration in Web Science: Instruments for Web Observatories

Marie Joan Kristine Gloria
Rensselaer Polytechnic Institute
Troy, NY
glorim@rpi.edu

Deborah L. McGuinness
Rensselaer Polytechnic Institute
Troy, NY
dlm@cs.rpi.edu

Joanne S. Luciano
Rensselaer Polytechnic Institute
Troy, NY
jluciano@rpi.edu

Qingpeng Zhang
Rensselaer Polytechnic Institute
Troy, NY
zhangq6@rpi.edu

ABSTRACT

The following contribution highlights selected work conducted by Rensselaer Polytechnic Institute's Web Science Research Center. (RPI WSRC). Specifically, it brings to light four different themed Web Observatories - Science Data, Health and Life Sciences, Open Government, and Social Spaces. Each of these observatories serves as a repository of data, tools, and methods that help answer complicated questions in each of these research areas. We present six case studies featuring tools and methods developed by RPI WSRC to aide in the exploration, discovery, and analysis of large data sets. These case studies along with our web observatory developments are aimed to increase our understanding of web science in general and to serve as test beds for our research.

Categories and Subject Descriptors

E.0 Data General

General Terms

Design, Human Factors

Keywords

Web Observatory, Linked Data, Methods, Semantic technologies

1.INTRODUCTION

As the Web matures, academics and researchers agree on the need to create, deploy, enable, and foster mechanisms and tools for its exploration and sustainability. The goal of a Web Observatory is to mobilize a research community that leverages the strengths of multiple disciplines, methodologies, and theoretical frameworks. At Rensselaer Polytechnic Institute's Tetherless World Constellation Web Science Research Center (RPI WSRC), our work addresses multiple facets of this goal including: the web's infrastructure, transdisciplinary data exploration and analysis, visualization, and social networks. As such, our observatories present both tools and methodologies that empower researchers to study the web and to make a difference in the world.

The RPI WSRC has four central themed observatories - science data, health and life sciences, open government, and social spaces. These four speak to a growing interest in each of these research areas and to where the collection and analysis of data has scaled significantly thanks to the Web. Our observatories include tools, collaborative processes, and methods that enable researchers to answer critical and complicated questions. To illustrate this, we present several case studies from each of these observatory themes. First, we introduce our most comprehensive observatory, the Science Data Observatory. Here we briefly discuss the SemantEco and Semantic Water Quality portal projects as exemplar projects; although, we have generated numerous science observatories. Second, the health and life sciences observatory includes our work on the Health and Human Services (HHS) Data Challenge. This effort is one of a number of health efforts, and it highlights a set of in-house developed tools that enabled the discovery of, access to, and integration of HHS's datasets. More importantly, our contribution exposes the power and efficiency of a semantics enriched toolkit and process. Third, we turn to our work in the open government space, which we demonstrate with our International Open Government Data Set (IOGDS). The IOGDS is a linked data application based on metadata "scraped" from hundreds of international dataset catalog websites publishing a rich variety of government data. Lastly, the RPI WSRC is developing the tools and methods to explore social spaces with the First Responder's Portal and the Twitter Network Observatory. Both enable the exploration of relationships and semantics in graph databases. In sharing our work, we hope to showcase how we can use the Web as a tool to study real world events; how semantics-enriched tools ease exploration within these Web observatories; and how we can now examine emerging communities on the Web.

2.SCIENCE DATA OBSERVATORY

The Semantic Ecology and Environmental Portal (SemantEco) facilitates collaborative work across multiple disciplines by providing support tools to help manage, analyze, visualize, and present large complex ecosystems. This semantically-enabled environmental monitoring framework uses a family of ontologies, some domain-independent aimed at facilitating monitoring; including for example, the notion of pollution events - when contaminant measurements are outside of appropriate ranges. SemantEco provides an OWL-based reasoning scheme and provenance-based facet generation to leverage query answering and data validation over the integrated data via SPARQL [2].

Specific topic portals then include domain dependent ontologies. For example, the Semantic Water Quality Portal (SemantAqua) includes ontologies concerning containments relevant to water pollution and relevant regulations. SemantAqua uses description logic reasoning along with its ontologies to detect water pollution

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

and facilities that violate water regulation. It preserves provenance metadata using the Proof Markup Language (PML) [7]. It is thus capable of providing detailed information about where the information came from and what inferences were done. It also integrates relevant health information so that it can connect health effects of exposure to high levels of contaminants.

Furthermore, the system also facilitates and encourages local community involvement. For example, if a community member is curious about contaminants that are outside of acceptable levels in their region, the portal can show on a map all of the polluted water areas or polluting facilities. Moreover, it can show which health effects are related to overexposure to those contaminants. The portal also enables users to comment on measurement sites and discuss information related to the captured data. It also provides links to federal and state agencies in order for community members to report water quality issues based on the current ZIP code. In addition, the SemantAqua portal provides trend tools to show how pollutant concentrations or quantities change over time through the display of changes in regulated characteristics (e.g. acidity, temperature) and pollutants (e.g. arsenic, lead) [2].

Both the SemantEco and SemantAqua portals exemplify the advantages of using semantic technology to address real world concerns. Moreover, these intelligent applications exploit the depth of which linked open government data can be useful, practical, and accessible to community members, researchers, and academics. The RPI WSRC science data observatory includes numerous other example portals, some of which provide more sophisticated reasoning on different kinds of data, which we believe, serve as exemplary approaches to future collaborative scientific research endeavors.

3.HEALTH AND LIFE SCIENCES OBSERVATORY

As the demand by both the private and public sectors grows for value added information services, we at RPI WSRC recognize the need for a specialized focus on the health and life sciences. Thus, the Health and Life Sciences Observatory includes examples like the Health and Human Services (HHS) data challenge. Issued in 2012, these seven complementary developer challenges encourage innovation in “data processing, discovery, analysis, and interpretation” [8]. In early 2013, the RPI TWC was awarded first place in the Metadata developer challenge for our contribution of in-house developed tools that enabled the discovery of, access to, and integration of the HHS’s 368 datasets. Specifically, the team set out to provide both an accessible, powerful toolkit and an automated process. Leveraging RPI TWC created tools including our csv2rfd4lod-converter, we developed a streamlined, replicable process to convert and enhance the metadata of the HHS datasets.

The csv2rfd4lod-automation tool is a custom application tailored for data ingest with an ultimate purpose of improving data access and integration. It is designed to aggregate and integrate multiple versions of many datasets with a variety of source organizations in an incremental and backward-compatible way. Next, the team looked to LODSPeaKr, another RPI TWC original framework, that helps create Linked Data applications and publish RDF data quickly and with minimal effort. LODSPeaKr provided a set of functionalities that improved the accessibility of the data for both humans and for machines. The LODSPeaKr framework provided content negotiation wherein every requested URI from the data was served as one of many serializations (i.e. RDF, HTML, etc) identified by the client’s request. Yet, LODSPeaKr’s most impressive feature is its simple, but powerful ability to create meaningful visualizations of the data. We highlight this for its two-fold value. First, these multiple data views enabled users to navigate and to explore the data. Second, the resulting series of visualizations and charts were helpful for humans to consume and better understand the data. Our overall goal was to illustrate the power, ease, and flexibility of these tools and to empower a wider range of data users in the future.

RPI’s challenge submission again illustrates the advantages and efficiencies of semantic-enabled technologies. Thanks to the tools that the RPI team created, the process of collecting, converting, standardizing, integrating, and analyzing a large amount of data can be achieved. Like SemantAqua or SemantEco, we encourage others to leverage these tools and methods in their own work as all of these tools are available on the RPI TWC website. For the health and life sciences, these types of tools can be especially important given the rise in available health related data. For example, we are currently applying this tool suite in another domain focusing on tissue samples, mutations, and diseases. Our health and life sciences observatory continues to evolve and is aimed at supporting a broader range of users by adding tools to support data mining and exploration by researchers outside of the Web Science discipline.

RPI WSRC further contributes to the development of Health Web Science as a sub-discipline of Web Science [3] to look for ways the Web can be used to help address the current health care crisis. This includes building, analyzing, and leveraging health web observatories. Given that the current model for delivering health care is unsustainable [1] and that the Web has emerged as a daily part of life for much of the western world, it may offer a route to finding solutions that help to mitigate rising health care costs and help make healthcare sustainable. Health Web Science studies the Web and technologies that use the Internet, their emergent properties, and how these can benefit society in the area of human health. RPI has co-organized two international Health Web Science Workshops, presented at Medicine 2.0 with a third workshop planned for May 2013. These workshops aim to foster investigation that will lead to better utilization of the web as well as to inform policy makers with evidence that will enable changes based on what is learned.

4.OPEN GOVERNMENT OBSERVATORY

Over the past several years, governments from across the world have made significant strides to incorporate transparency practices. As a result, the web now hosts a plethora of open government data sets ranging from regulatory standards to bird migration patterns. The vastness and incongruence of such data can make it difficult to work with and may be intimidating for many researchers. For the last few years, the RPI TWC has worked to mitigate these concerns by developing a linked open data portal [4] and tools like the International Open Government Data Set (IOGDS) application. IOGDS is a linked data application that takes metadata from hundreds of international dataset catalog websites that publish a rich variety of government data. The metadata is then automatically converted to RDF linked data and re-published via the RPI TWC LOGD SPARQL endpoint where it is made available for download. IOGDS currently searches over 1 million datasets and over 197 catalogs from around the world, and the number keeps growing.

Like SemantAqua or the HHS Developer Challenge, the IOGDS application serves as yet another example of the power of linked data and semantic technologies. The challenges of collecting and searching datasets from various different sources are eliminated. Instead, researchers may visit the IOGDS demo page to search all 197 catalogs on a faceted S2S browser developed by RPI TWC. The S2S framework leverages the machine-readable semantics of data, services and user interface components while automating various tasks in UI development for the search interface [6].

5.SOCIAL SPACES OBSERVATORY

The rise of social media and its influence on our daily contemporary lives has many clamoring to unlock its potential for answering some of our most complicated human behavioral questions. At RPI WSRC, we too recognize this phenomena and are excited to introduce the Social Spaces Observatory; the latest addition to our set of observatories. The following section highlights two prototypes - the NIST/RPI TWC First Responders Network Observatory and the Twitter Network Observatory.

Our National Institute of Standards and Technology funded effort began in 2012. We were asked to design a first responder requirements gathering methodology that leverages social media. Further the effort includes designing and implementing prototype tools that facilitate the methodology [5]. The effort is examining the current state of collecting and synthesizing responder requirements, and is providing an assessment of the effectiveness of that process, along with an evaluation of existing candidate platforms for use within this community, and the production of a solutions roadmap [5]. We have developed a number of tools, initially focused on Twitter, to help gather relevant contact points related to first responder topics and tools to help analyze their networks.

The Twitter Network Observatory project, like the First Responders Network, also seeks to build a semantic-enabled platform that aggregates, stores, and analyzes Twitter data. The data is converted to RDF linked data and re-published using a TWC LOGD SPARQL endpoint. By doing this, researchers can explore the relationships of people and semantics in 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 other toolkits.

Once again, this observatory demonstrates RPI WSRC's focus on designing semantics-enriched technologies that enable the collection, integration, discovery, and analysis of large data sets. Furthermore, as we address in the First Responders Network, there is a need to develop methodologies in order to better leverage these types of online community platforms. However, we recognize that the social space observatory provides a unique challenge unlike those associated with our other three observatories. For this particular space, one next step is to move beyond just the data collection, integration, and initial analysis. As we demonstrate with the Twitter Network Observatory, it is in implementing features that allow for exploration of relationships over the graph that may be of most use.

6.CONCLUSION

We presented four Web Observatory themes hosted by the RPI WSRC. We identified these themes as key topics of interest and research both within our lab and in the general community. Within each observatory, we discussed specific examples such as SemantEco and the Twitter Network Observatory. These examples speak to our continued commitment to designing, building, and evaluating semantically-enriched tools to aide in the aggregation, integration and analysis of large data sets. Moreover, the lab has developed several tested methodologies that work in-step with these tools to ensure a seamless collaborative experience. Lastly, as we note in the social spaces observatory, we recognize the growing importance of this phenomena and its unique burden of serving as a mirror of ourselves. As such, there are additional challenges in developing the correct tools and methods for this space, which we are just now beginning to explore. And, while we believe that Web observatories serve as exemplary collections of tools, methods, and case studies for others to leverage, we must also consider future legal, ethical, and social implications.

7.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.

8.REFERENCES

[1] Burrill Report. (2012). "Market Turmoil Overshadows Biotech Successes in 2011". The Burrill Report. Volume 3, Issue, 3, February 2012.

[2] Challenge. (n.d.). "Apps for the Environment: SemantAqua Water Portal". Challenge.gov Website.

[3] Cummings, G., Luciano, J., Baker, C.J.O. & Cambria, E. (2012). "Health Web Science: Second Web Science Health Workshop." Workshop at Web Science Conference 2012. Evanston, IL.

[4] Ding, L., DiFranzo, D., Graves, A., Lebo, T., Erickson, J.S., . . .Hendler, J. "TWC LOGD: A portal for linked open government data ecosystems". *Journal of Web Semantic*s. Volume 9, Issue 3, September 2011, Pages 325-333.

[5] McGuinness, D. and Erickson, J., (2012). "First Responders Requirements Methodology". TWC Website. http://tw.rpi.edu/web/project/FirstResponders

[6] Rozell, E., Fox, P., Maffei, A., Zednik, S. (2011), A Framework for Earth Science Search Interface Development, Abstract EGU2011-13413 presented at General Assembly 2011, EGU, Vienna, Austria, 03-08 Apr (slides) - See more at: http://tw.rpi.edu/web/project/SeSF/workinggroups/S2Sold#sthash.YNoM5bqV.dpuf

[7] Wang, P., Fu, L., Patton, E.W., McGuinness, D.L., Dein, J., and Bristol, S. 2012. Towards Semantically-enabled Exploration and Analysis of Environmental Ecosystems. In

Proceedings of 8th IEEE International Conference on eScience (October 8-12 2012, Chicago, IL).

[8] 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/