Towards Semantically-Enabled Next Generation Community Health Information Portals: The PopSciGrid Pilot

Deborah L. McGuinness
Li Ding
Tim Lebo
James P. McCusker

Rensselaer Polytechnic Institute
Troy, NY
dlm@cs.rpi.edu, lebot@rpi.edu,
dingl@cs.rpi.edu, mccusj@rpi.edu

Abdul R. Shaikh
Glen D. Morgan
Gordon Willis
Richard P. Moser
Zaria Tatalovich
Bradford W. Hesse

National Cancer Institute
Rockville, MD
{shaikhab, moserr, gmorgan,
tatalovichzp, willisg, hesseb}@mail.nih.gov

Noshir Contractor
Northwestern University
Evanston, IL
nosh@northwestern.edu

Paul Courtney
SAIC-Frederick, Inc
Frederick, MD
courtneypk@mail.nih.gov

Abstract

We describe an approach to developing next generation health information portals. This prototype portal was developed to address two complementary goals (1) design and create a site where people can explore potential relationships between selected health-related behaviors, policies, and demographic data (2) explore semantic web technologies and linked data as enabling technologies for next generation health informatics portals. Our multidisciplinary team includes population and behavioral scientists, social network scientists, statisticians, and computer scientists focused on creating innovative proof of concept applications that integrate complex health data in understandable and usable ways. Our semantic-web based framework allowed us to design exemplar community health portal applications, with an initial focus on tobacco-related health data such as smoking prevalence and tobacco policies (taxation and smoking bans). We describe our approach, two semantically-enabled tobacco-related applications, and discuss how this approach can be used in a broad spectrum of community health applications.

1. Introduction

As policy makers, researchers, and consumers strive to find ways to improve overall health, reduce disease, and simultaneously manage health-related costs, many look to informatics for guidance. Our work aims to create information portals that contain relevant information that users find understandable and trustworthy enough to take action. Ultimately, the goal is to enable people to make informed decisions related to personal and population health. Our aim is not just to allow users to find and access information but also to allow them to utilize it as they form and investigate additional hypotheses. We are creating tools and infrastructure based on semantic technologies to support aggregation, integration, and presentation of information. We believe this PopSciGrid pilot effort provides a first step in demonstrating what can be done with such tools and infrastructure and in this paper we intend to exemplify some of the benefits and challenges of using semantic technologies in one example next generation community health portal.

The PopSciGrid Community Health Portal was developed with a multidisciplinary team. The diverse team was formed so that we could represent differing needs and points of view and so that we could synthesize the various requirements from many different perspectives. The team consists of semantic web, web science, and social science academic researchers, government population science and cancer control and prevention researchers, health data analysts, and related contractors. Some initial motivating health-related questions include: "*Can we integrate health policy information with health and behavior data in understandable and useful ways?"* as well as *"Can we help policy-makers form and investigate hypotheses about which policies may be correlated with behavior changes?"* Some initial motivating semantic web-related questions include: *Can we leverage semantic technologies to support health data integration?"* as well as *"Can we leverage linked data resources to allow a broad range of users to more easily build health information portals?"*

PopSciGrid explores the intersection of health behavior, policy, and demographic data using a Linked Data powered platform. With the central importance of tobacco control for public health, we chose an initial domain related to tobacco policies, smoking prevalence, and related demographics for the initial proof of concept pilot focus. The PopSciGrid demos serve both as demonstrations of potential interactions between tobacco policy and tobacco-related behaviors[1] and also as an operational specification of how to build a semantically-enabled community health portal. As such, the primary audience includes those interested in using health information portals – initially those interested in community health information portals related to tobacco information. For example, policy makers, tobacco researchers, and interested lay people could benefit. A second, complementary goal of the work – as an operational specification – targets an audience interested in the tools, infrastructure, and design required to build such portals. Initially, this would involve implementers literate in semantic technology and linked data. In the rest of this paper, we will describe the tobacco-centered health information portal and the semantic technologies behind it. We will then identify benefits from the semantic approach and discuss some current and future directions.

2. Background

The tobacco and population health researchers in our team identified data sets relevant to tobacco-related health data. Originally, the small team identified variables pertinent to smoking prevalence from the National Health Interview Survey (NHIS[1]) and the Health Information National Trends Survey (HINTS[2]), two publicly available health survey data sets. The initial portal design [11] employed tools and infrastructure from the Cancer Biomedical Informatics Grid (caBIG®), a government sponsored research platform to standardize and share cancer data, in order to integrate data from a portion of the data available. The interface was designed to facilitate basic exploration of the data contained in each dataset with summary statistics dynamically computed based on filters applied, and a map display of the data geo-coded at the state and regional levels. The feedback received from population and tobacco researchers who viewed demonstrations of this portal indicated that there was significant promise of enabling the exploration of data gathered across disparate datasets in order to provide insight into health-related questions.

These early successes motivated the need for more automated approaches to data input integration and visualization. In 2009, experts in Semantic Technology and Linked Data were brought onto the team to incorporate design techniques that leverage the potential from the open linked government data movement. In particular, they looked to leverage the platforms and tools built around the United States' data.gov content such as the Rensselaer Tetherless World Constellation Linked Open Government Data (LOGD) platform [3] or the Southhampton open data efforts[3][4]. Using these new data integration approaches and tools, the extended team considered additional datasets from impacteen.org (including a data summary of state-level smoking prevalence) and smoke-free policy coverage datasets curated by the National Cancer Institute (NCI). The current infrastructure design also uses a number of Linked Data tools for conversion of data, integration, provenance representation, and visualization. The resulting, more general, portal [6, 2] with expanded domain content and semantically-enabled applications is available on the web[5].

3. Community Health Portal

The current PopSciGrid pilot project aims to provide data in an interactive, data-driven interface that enables users to easily explore contextual factors that may impact smoking prevalence. For example, users may want to look for correlations between price per cigarette pack or tax per pack and smoking prevalence. Users may also want to view data associated with smoking bans in bars, restaurants, and workplaces and then they may wish to look for potential relationships between these bans and smoking prevalence, utilizing data in user-friendly ways that are not possible with the raw data in its original form.

Figures 1 and 2 demonstrate one visualization approach. The visualization uses the motion chart component of the Google visualization toolkit[6]. The axes in the graph are reconfigurable. In Figure 1, we display **tax per pack** on the *y* axis and a **measure of smoking bans**[7] on the *x* axis. The demonstration

[1] http://www.cdc.gov/nchs/nhis.htm
[2] http://hints.cancer.gov

[3] http://data.southampton.ac.uk
[4] http://opendatamap.ecs.soton.ac.uk/
[5] http://logd.tw.rpi.edu/project/popscigrid
[6] http://code.google.com/apis/chart/interactive/docs/gallery.html
[7] The ban policy data in this demonstration is averaged across all three types of policies (restaurant, workplace, and bars).

allows someone to "play" the data over time and view the smoking prevalence data as it changes through time. Each state is displayed with a circle representing the relative size of smoking prevalence of its citizens. The prevalence is also broken down into high (blue), medium (green), and low (yellow) prevalence. Figure 1 shows the initial starting point for 1991, when no smoking policy bans were in effect (in workplaces, bars, and restaurants) and taxes per pack were relatively low.

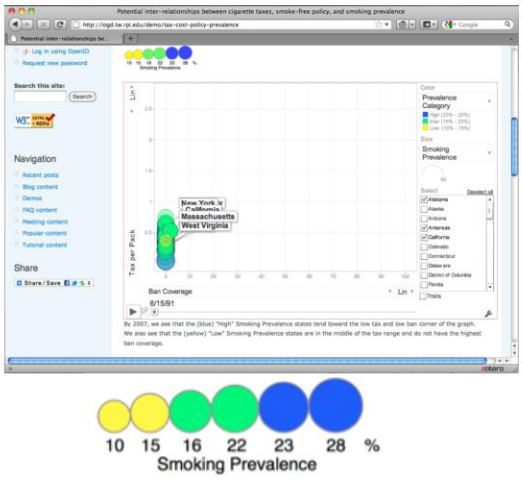

Figure 1. Taxation, Policy, and Prevalence in 1991.

Figure 2. Taxation, Policy, and Prevalence by 2007.

Figure 2 shows the result of running the demonstration motion chart over time. If one displays this interactively, one can see that some states like California took an early lead in putting smoking bans in place. We can also see that some states like New Jersey have put in place broad smoking ban policies as well as high taxation policies. When viewing a run of the demo, it also becomes easy to see the states that tried more or less aggressive policies and when they did so in comparison to the other states. In both Figures 1 and 2, only the states that have been selected using the checkboxes on the bottom right are labeled in the graph, allowing users to focus more easily on individual states and help to reduce some visual clutter. The interactive version of this demonstration is available on the web[8].

Figure 3 uses another Google visualization toolkit component and shows a map of smoking prevalence *by state* in a year that can be chosen from a pull down. This image shows data from 2007. The lighter colored states, such as Utah and California, have the lowest smoking prevalence and the darker states have higher prevalence numbers. The graphs on the right are configurable to allow the user to select any state and view data related to it. In this demonstration, ban policies in workplaces, bars, and restaurants are shown individually in the bar chart, so for example, users can view when states put in workplace smoking bans different times from when they put in restaurant or bar smoking bans. One can also view cigarette price and tax over time. This second demonstration is also available on the web[9].

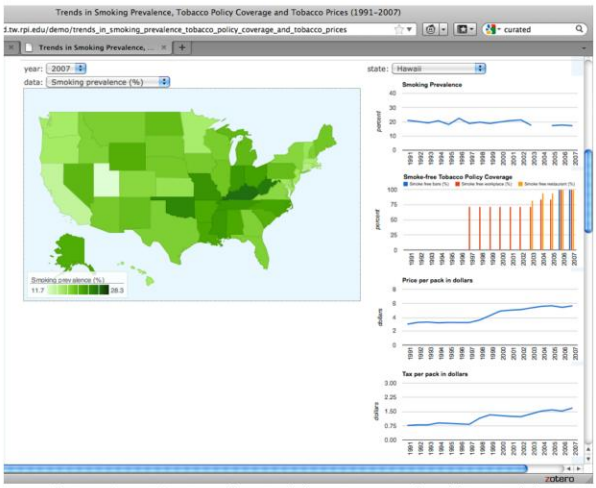

Figure 3. US overview with per-state details graphing prevalence, ban policies, price, and tax.

The two demonstrations shown in Figures 1-3 are generated from the same data but employ two different visualization styles. Each may be better for different users or for the same user with different goals. For example, one can see from the second demonstration in Figure 3 that Hawaii introduced some smoke-free policies in workplaces as early as 1997, but did not put

[8]http://logd.tw.rpi.edu/demo/tax-cost-policy-prevalence
[9]http://logd.tw.rpi.edu/demo/trends_in_smoking_prevalence_tobacco_policy_coverage_and_tobacco_prices

similar policies in bars until 2006. However it also shows that when the smoke free policies went into effect in 2006, the policies provided 100% coverage across the state. The visualization in Figure 3 may be more useful for users who are interested in drilling down into individual policy bans. The visualization in Figure 2 may be more useful for users who are interested in getting a sense of which states started to experiment with policies first and which states followed suit later. Note that this information could also be obtained from the visualization in Figure 3 (as long as one was willing to take an average figure for policy bans) but it would require clicking on each individual state to gather the policy data rather than a single click in Figure 2 that "played" the data over time in one nation-wide run of the data.

In addition to the two described here, numerous other demonstrations were also generated from the same underlying data – some using graduate students and some undergraduates. We found it easy for people somewhat familiar with visualization toolkits, such as the Google Visualization toolkit, to take this data and explore different ways of viewing hypotheses exposing possible relations between multiple parameters. This paper highlights only the two featured demonstrations but some of the other investigation directions are listed on the public project page[10] under experimental status demonstrations, and others are also available upon request.[11]

4. Methods

Summary data (i.e., smoking prevalence by state) were gathered from the ImpacTeen State Level Tobacco Control Policy and Prevalence Database [4] covering state tobacco control policy and prevalence data. Data were also gathered from the National Cancer Institute covering changes in tobacco ban and taxation policy based on the Chronological Table of U.S. Population Protected by 100% Smokefree State or Local Laws available from ANRF[12] from 1990 to 2007. A detailed description of the data gathering and analysis effort is in process in [12].

Because all source data were in free-form tabular format (Excel and CSV files), the *csv2rdf4lod*[13] conversion software was used to automate the creation of Resource Description Framework (RDF) representations that were hosted using Linked Data principles. While *csv2rdf4lod* performs a role similar to the RDF export extension available for Google Refine[14], it uses a declarative mapping description that permits others to automatically reproduce the conversion. The Google Refine RDF extension, on the other hand, is designed for interactive mapping and cannot be easily shared, reproduced, or modified by subsequent investigators. Thus, *csv2rdf4lod* is more effective for use in large scale, repeatable, and distributed data integration [5]. Further, the n-ary relations expressed by the data sources' statistical measures are not supported by the simplifying assumptions of Google Refine's RDF export extension, while both forms of interpretation are uniformly handled by *csv2rdf4lod*[15]. The *csv2rdf4lod* converter also tracks provenance: i.e., it records the URLs retrieved from the original data providers, when the URLs were retrieved, when and how the original data files were converted to RDF, where dump files are available (for subsequent analysis), and when these RDF dump files were loaded into the triple store. The conversion configurations are also available to allow others to inspect or repeat any processing applied after retrieving the original data from the authoritative source. Provenance information is described using the PML Provenance Interlingua [7].

Capturing the provenance along with the raw data allows applications to query data elements according to how they were derived and enables support for interfaces and applications to provide provenance-based filtering (such as only obtaining data from particular sources, or those involved in particular kinds of analyses). Our demonstration applications allow users to have increasing access to the provenance data, thereby allowing them to obtain information that may be relevant to their decisions to use and act on the data and conclusions from the portal. This workflow is illustrated in Figure 4.

[10] http://logd.tw.rpi.edu/project/popscigrid

[11] Only demonstrations that were vetted by our NCI colleagues are included in the featured demonstration listing but other demonstrations were partially generated as we experimented with possible data visualization and mash-up opportunities. We also welcome input concerning other potential desired parameters to expose in related visualizations.

[12] http://www.no-smoke.org/pdf/EffectivePopulationList.pdf

[13] http://logd.tw.rpi.edu/technology/csv2rdf4lod

[14] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/

[15] https://github.com/timrdf/csv2rdf4lod-automation/wiki/Converting-with-cell-based-subjects

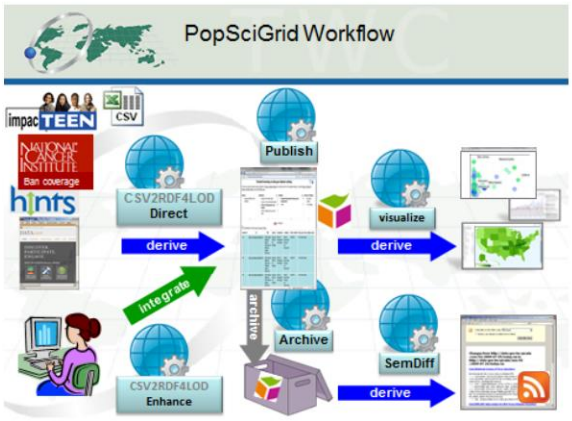

Figure 4. Data Conversion to Visualization Workflow

In the converted RDF data, variables (such as smoking prevalence, cigarette taxation, and smoke-free policy coverage) and dimensional parameters (such as state and year) are explicitly represented, which makes dataset integration relatively straightforward because they share common data entities. For example, we routinely join datasets by state (or geographic region) and by time periods (such as year or month). These data and its provenance were stored in an RDF database or 'triple store'. While we use OpenLink Virtuoso[16], any triple store conforming to the SPARQL specification would be sufficient. The Javascript visualizations query the triple store using SPARQL, retrieve results in JSON, and populate a variety of Google Visualization APIs. The featured demonstrations use Google visualizations, but the web standards used by the system permit other mechanisms to access and apply alternative analyses and visualizations.

We found that using RDF to build different visualizations is useful because it explicitly connects the disparate data sets. When the data are more consistently connected, queries are easily developed, analysis is easily performed, and results can be visualized easily using a variety of visualization tools aimed at taking RDF input. As a simple example, states were mentioned using a variety of identifiers (i.e., "ID", "Idaho") that would make it difficult to query without custom application logic. Addressing these issues at an earlier stage and unifying these references to a common semantic web Uniform Resource Identifier, such as http://logd.tw.rpi.edu/id/us/state/Idaho , allows application developers to avoid special case logic in each visualization. The explicit connections among the data sets can also facilitate exploration and recommendations for queries, analyses, and visualizations to develop in subsequent investigations. Using RDF to integrate disparate datasets has advantages over proprietary integrations from Business Intelligence Tools because RDF is designed for web-ready publication (as Linked Data) that can be incrementally augmented by third parties from around the world without *a priori* coordination. This open framework allows the value of the individual PopSciGrid efforts to accumulate as others find, augment, and use our open, published data.

5. Discussion

Using the flexible RDF-based triple format, we found it relatively simple to take the initial manual demonstration system and replicate it in a more automated, extensible, and scalable manner. Even though these demonstrations only display a small set of the parameters in our triple store, it is straightforward for users to explore relationships between any of the parameters. For example, we have explored relationships between education level, job status (employed or unemployed), self-reported depression levels, and smoking prevalence. By using the automated tools, we have a consistent and relatively complete encoding of provenance that we can expose in our demonstrations and search interfaces. We have previously and continue to investigate faceted search interfaces that allow users to constrain values for a potentially large number of parameters as they search through and filter the data. Sometimes the search constraints that one wants to use relate to a raw data value. For example, one may want to search for states whose smoking prevalence after a certain time was above a particular percent – for example find all states after 2001 whose smoking prevalence was high – or above 23%. The constraints may also be relative in nature. For example those looking for states going in the right direction with respect to smoking, one may want to search for states whose smoking prevalence has fallen for the last 5 years. Similarly people may look for states with consistent smoking problems such as those that have been among the highest five prevalence values in the last five years. Sometimes the constraints on search may be related to the meta data. For example, one may want to retrieve only data that was from a particular trusted source or was more recent than a particular date.

As a result of converting the data into a standard triple format, there is a large variety of visualizations

[16] http://virtuoso.openlinksw.com/

that can be generated from our data. These visualizations are relatively easy to generate from triplified data, so it is simple to generate a number of alternative views of the data (such as the 2 views displayed in Figures 1-3). One benefit of the variety of visualizations was that it allowed for identification of some anomalies that became obvious when viewing data in specific forms (e.g., some visualizations clearly identified values above 100% which led us to discover and report some rounding errors in the government's smoking policy ban data). Starting with this error, we were able to utilize the recorded provenance trace as the basis of communication with government data curators to investigate and fix issues in the original data. We have found that this flexibility in format and content to be particularly useful in working in a broad multidisciplinary team.

We also have begun to reach out to other efforts that have domain overlap. As mentioned above, the first iteration of this portal used caBIG® tools and technology on the back end. While the newer PopSciGrid demonstration used a semantic web and linked open data backend infrastructure instead, it would be more valuable if we could leverage resources from both communities. To fill this gap, we developed a tool called swBIG that establishes semantic web identifiers (URIs) for the entities available within caBIG® services and acts as a bridge by responding to requests from either infrastructure. This leverages the development of both communities while at the same time exposes each data source in alternative environments. Figure 5 illustrates this connection from caBIG® to the PopSciGrid component of the semantic web. We show in McCusker *et al.* [8] that accession of data through the swBIG proxy provides data tied to well curated controlled vocabularies such as the NCI Thesaurus, which is not directly available from the originating grid services. Although this version of PopSciGrid did not utilize caBIG® directly, the promise of the two demonstrations we discuss in this paper has shown that it is worth investigating how they can be extended to use swBIG proxy to access caBIG®.

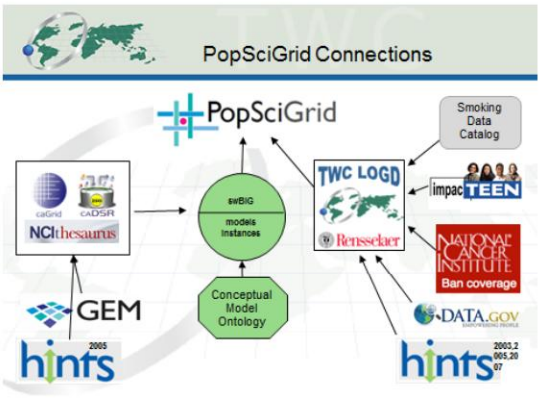

Figure 5: The swBIG service bridges PopSciGrid to other resources including NCI's caBIG® infrastructure.

6. Directions

We have been encouraged by the directions of our work on community health information portals. Our experience is that the automated tools make it easier to create new demonstrations from multiple perspectives by a broader range of people than previously possible. For example, computer scientists who had little knowledge of tobacco issues could create starting points for tobacco researchers to review relatively simply. The conversion tools make it easier to ingest the data from a variety of disparate sources and put it in a standard format while maintaining provenance related to where the data came from and what transformations have been applied. Once in triple form, there are a large number of visualization tools that can be used to present data. The semantic annotations make it easier to understand what the data represents and thus makes it easier for others to mash the data up with other datasets and create meaningfully integrated demonstrations.

The pilot project demonstrations provides some initial views of what can be done with the framework, thus allowing others with data or with data use cases to envision how the framework might be used in their domains. For example, after we demonstrated the PopSciGrid portal at a recent talk at NIH, another researcher approached us who had data about policies related to physical education time requirements and nutrition policies in elementary, junior, and senior high schools from the Classification of Laws Associated with School Students (CLASS)[17]. We began to explore the potential relationships between policies related to physical education requirements and nutrition in

[17] http://class.cancer.gov/About.aspx

schools, using visualizations to explore associations with between policy and health outcomes. Following the same workflow shown in Figure 4, we quickly converted the data into triple form. Further, an undergraduate who had not been working with these tools generated some initial demonstrations similar to the demonstration shown in Figure 3 in just a few hours. While that demonstration is not as robust nor as well investigated as the current PopSciGrid, it was instructive for showing that the approach can be used in other health domains and also that someone with relatively little training can quickly generate initial prototypes. The relative ease with which early prototypes can be created provides an additional set of tools for communication across disciplines. In the CLASS project situation, it allowed us to not just describe in words what we envisioned for the CLASS data, but also allowed us to show in an inexpensive demonstration what we might do. It helped us interact with domain experts when we could show them our general framework applied to their data and then allowed us to get more of their help in determining requirements and high value features. It also helped us brainstorm with domain experts to identify other related data sets or other parameters or policies that would be valuable to explore. For example, one obvious extension to the demo related to the CLASS data is to obtain data on childhood obesity and investigate potential relationships between the policies in regions and their obesity rates.

Our work on health information portals is still in its early stages. These demonstrations are aimed at exploration and hypothesis formation for further investigation. We intend to expand our initial demonstration to include and explore relationships between smoking prevalence and other data, including additional demographics, risk factors and potentially relevant health statistics such as lung cancer incidence and mortality data for our tobacco demonstrations. We are also exploring additional visualizations. In addition, we plan to use the same platform to generate demonstrations in other key health-related behaviors. To evaluate the effectiveness of these technological advances, we are exploring ways to engage a broader community and one initial step in our plan is to conduct usability testing with members of relevant communities

We are also exploring different levels of granularity. While most of our demonstrations use state level data, we are investigating the benefits of zip code level granularity. We expect more flexibility for mash ups against additional demographic data and see the potential to allow lay people as well as policy makers to investigate possible trends in policies at different geographic levels. Additional granularity also gives more flexibility for the types of integrated demonstrations we may provide. For example, we initiated a portal called HealthScore[18], which takes a zip code as input and displays health data relevant to that zip code. It allows a user to get a sense of the relative health of a region by looking at values and their relationship to the mean for parameters such as longevity, air quality, and live births.

We are investigating options for additional usage of and enhancements toour work with provenance. For example, work continues on our *csv2rdf4lod* converter and on additional examples of tools and services that leverage more of the provenance we are currently maintaining. In another application of our framework focused on environmental informatics [13], our interfaces include additional support for providing more provenance-aware search and additional provenance follow up support. For example, when users find information that appears surprising, they may want to further investigate issues concerning the source(s) of the data and any transformations or interpretations that are relevant. It is an open social question as well as an open technical question concerning what provenance and background information is most appropriate to show when users would like more information about how to interpret what they are viewing in a demonstration. Our current strategy is to generate a number of interface and technical strategies and then to gather some user feedback to help develop and refine provenance services and presentation strategies.

We are also pursuing annotation schemes that allow users to interact with the data and leave annotations on it [9]. For example, users may see data embedded in an application that they believe is wrong or refuted by other data. Our annotation work allows users to annotate data and either ask questions or leave comments on it. We have also begun an effort on a semantic application framework that allows programmers to understand much less about the overall structure of the hardware because they have tools that do some amount of the integration and “plug and play” by themselves [10].

There are a number of issues that remain interesting and open and potentially go far beyond just technical considerations. For example, while the current demonstrations provide ways to investigate hypotheses about policies, it is not clear how best to engage policy makers to use the system to pose, refine,

[18] http://logd.tw.rpi.edu/demo/healthscore

and test potential policy hypotheses. For example, the system is currently configured to support review of tax, price, and particular bans relative to smoking prevalence. It could easily be configured to look at these issues along with other parameters that we already have data for (such as education level, age, gender, etc), or parameters for which data may need to be obtained.

Currently we believe that the system holds the most potential when reviewed by teams of people – for example tobacco researchers, policy makers, and technologists – in order to identify the most promising additions. Our goal however is to provide additional tools and interfaces that are aimed at specific audiences so that for example, policy makers could review and reconfigure the demonstrations in real time to investigate their favorite topics. The current demonstrations are easily configurable for computer scientists but often these people do not have enough tobacco and policy insight to create compelling demonstrations of *meaningful* patterns, trends, correlations in the data. With input from domain- and policy- literate audiences we believe we can provide additional tooling that could meet at least some of their requests so that they could directly generate new meaningful demonstrations from augmented interfaces without computer scientists in the loop. Further there is potential out of these kinds of systems for people to review a demo such as PopSciGrid and then post on a social network perhaps a question such as "*Do tobacco bans really reduce smoking prevalence? This visualization seems to imply it does. What other related data supports or refutes this hypothesis.?*"

One path for demonstration systems such as PopSciGrid is for them to provide exploration tools and community focal points for helping to initiate discussions that are grounded in reputable, trustworthy data. These discussions may engage communities to help identify promising hypotheses and even provide some initial analyses of additional relevant data. This could entail envisioning the PopSciGrid as a multidimensional network that enables computation and collaboration among a diverse set of nodes including individuals, datasets, concepts, and analytic tools [1]. We can envision community portals that have as key components interactive demonstrations such as PopSciGrid. These portals then can help to engage broader audiences in discussions around policies and possible causal relationships that may help empower and energize communities to initiate new programs and policies. One key to this approach is to include interactive and reconfigurable demonstrations such as PopSciGrid that serve as windows onto large data portals.

7. Conclusion

In this paper, we have introduced our work on semantically-enabled community health portals. We have described our design that allowed us to ingest, integrate, annotate, and visualize health policy and related health data. We have described a particular pilot community health portal that explores how tobacco-related policies and behaviors can be integrated, visualized, and communicated to empower communities and support new avenues of research and policy for initially aimed at cancer prevention and control. We identified and demonstrated by example some of the value of creating applications utilizing a semantic web infrastructure. We further hypothesized that this infrastructure can be reused in many community health portal settings and provided some initial evidence supporting this hypothesis.

8. Acknowledgements

We sincerely thank Erik M. Augustson, Kelly Blake, Hugh Devlin, Lila Finney, Yvonne Hunt, Amy Sanders, Yun Huang, and York Yao for their assistance throughout the development of the work presented in this paper.

9. References

[1] N. Contractor, P. Monge. And P Leonardi. Multidimensional networks and the dynamics of sociomateriality: Bringing technology inside the network. *International Journal of Communication*, 5, 1-20, 2011.

[2] P. Courtney, A. R. Shaikh, N. Contractor, D.L. McGuinness, L. Ding, E. M. Augustson, K. Blake, G. D. Morgan, R. Moser, G. Willis, and B. W. Hesse, Consumer Health Portal: An Informatics Tool for Translation and Visualization of Complex, Evidence-Based Population Health Data for Cancer Prevention and Control, In 138th APHA Annual Meeting, 2010.

[3] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, D.L. McGuinness, and J. Hendler. TWC data-gov corpus: incrementally generating linked government data from data.gov. In 19th Intl World Wide Web Conference. 2010.

[4] G. A. Giovino, F. J. Chaloupka, A. M. Hartman, K. Gerlach Joyce, J. Chriqui, C. T. Orleans, K. Wende, C. Tworek, D. Barker, J. T. Gibson, J. Yang, J. Hinkel, K. M. Cummings, A. Hyland, B. Fix, M. Paloma, and M. Larkin. Cigarette Smoking Prevalence and Policies in

the 50 States: An Era of Change. – The Robert Wood Johnson Foundation ImpacTeen Tobacco Chart Book. Buffalo, NY: University at Buffalo, State University of New York, 2009. http://impacteen.org/statetobaccodata/chartbook_final060409.pdf

[5] T. Lebo, J. S. Erickson, L. Ding, A. Graves, G. T. Williams, D. DiFranzo, X. Li, J. Michaelis, J. G. Zheng, J. Flores, J., Z. Shangguan, D. L. McGuinness, and J. Hendler. Producing and using linked open government data in the twc logd portal. In Wood, D., ed., Linking Government Data. New York, NY: Springer. (in press) 2011.

[6] D. L. McGuinness, A. R. Shaikh, R. Moser, B. W. Hesse, G. D. Morgan, E. M. Augustson, Y. Hunt, Z. Tatalovich, G. Willis, K. Blake, P. Courtney, L. Finney, A. Sanders, L. Ding, T. Lebo, J. McCusker, N. Contractor, Y. Huang, Y. Yao, and H. Devlin. A Semantically-enabled Community Health Portal for Cancer Prevention and Control. In Proc. Third International Web Science Conference, Koblenz, Germany, June 15-17, 2011.

[7] D. L. McGuinness, L. Ding, P. Pinheiro da Silva, and C. Chang. PML 2: A Modular Explanation Interlingua. In Proc. Of the AAAI '07 Workshop on Explanation Aware Computing, July 2007. ftp://ftp.ksl.stanford.edu/pub/KSL_Reports/KSL-07-07.pdf

[8] J. P. McCusker, J. S. Luciano, and D. L. McGuinness. 2011. Towards an Ontology for Conceptual Modeling, In Proc. of the International Conference on Biomedical Ontology, 2011. http://tw.rpi.edu/web/doc/towardscmo

[9] J. Michaelis, S. Zednik, P. West, P. Fox, and D.L. McGuinness. 2010. Extending eScience Provenance with User-Submitted Semantic Annotations, Abstract IN43C-08. In Proceedings of AGU Fall Meeting 2010 (December 13-17 2010, San Francisco, CA).

[10] E. Patton, D. DiFranzo, and D.L. McGuinness. . SAF: A Provenance-tracking Framework for Interoperable Semantic Applications. In: 3rd International Provenance and Annotation Workshop. (June 24, 2010).

[11] A. R. Shaikh, N. Contractor, R. Moser, G. D. Morgan, P. K. Courtney, E. M. Augustson, A. M. Pilsner, and B. W. Hesse. PopSciGrid: Using cyberinfrastructure to enable data harmonization, collaboration, and advanced computation of nationally representative behavioral, demographic, and economic data. In 137th APHA Annual Meeting, (2009).

[12] Z. Tatalovich, Y. Hunt, S. Marcus, N. Howlader, and A. Mariotto. Geographic Patterns of Local Tobacco Control Ordinances and Cigarette Consumption in the US. In process.

[13] P. Wang, J. Zheng, L. Fu, L., E. Patton, T. Lebo, L. Ding, Q. Liu, J.S. Luciano, and D.L. McGuinness, 2011. TWC-SWQP: A Semantic Portal for Next Generation Monitoring Systems. In Proceedings of 10th International Semantic Web Conference (October 23-27 2011, Bonn, Germany).