Towards Next Generation Health Data Exploration: A Data Cube-based Investigation into Population Statistics for Tobacco

James P. McCusker*, Deborah L. McGuinness*, Jeongmin Lee*, Chavon Thomas†,
Paul Courtney‡, Zaria Tatalovich§, Noshir Contractor¶, Glen Morgan§, and Abdul Shaikh§
*Rensselaer Polytechnic Institute, Troy, USA
Email: {mccusj, dlm}@cs.rpi.edu, leej35@rpi.edu
†Hobart and William Smith Colleges, Geneva, NY, USA
Email: chavon.thomas@hws.edu
‡Dana-Farber Cancer Institute, Boston, MA, USA
Email: Paul_Courtney@dfci.harvard.edu
§National Cancer Institute, Bethesda, MD, USA
Email: {tatalovichzp,gmorgan,shaikhab}@mail.nih.gov
¶Northwestern University, Evanston, IL, USA
Email: ncontractor@gmail.com

***Abstract*—Increasingly, experts and interested laypeople are turning to the explosion of online data to form and explore hypotheses about relationships between public health intervention strategies and their possible impacts. We have engaged in a multi-year collaboration to use and design semantic techniques and tools to support the current and next generation of these explorations. We introduce a tool, qb.js, to enable access to multidimensional statistical data in ways that allow non-specialists to explore and create specific visualizations of that data. We focus on explorations of health data - in particular aimed at helping to support the formation and analysis of hypotheses about public health intervention strategies and their correlation with health-related behavior changes. We used qb.js to formulate and explore the hypothesis that youth tobacco access laws have consistent, measurable impacts on the rate of change in cigarette smoking among high school students over time. While focused in this instance on one particular intervention strategy (i.e., limiting youth access to tobacco), this analytics platform may be used for a wide range of correlational analyses. To address this hypothesis, we converted population science data on tobacco-related policy and behavior from ImpacTeen to a Resource Description framework (RDF) representation that was annotated with the RDF Data Cube vocabulary. A Semantic Data Dictionary enabled mapping between the original datasets and the RDF representation. This allowed for the creation and publication of data visualizations using qb.js. The RDF Data Cube representation made it possible to discover a significant downward effect from the introduction of nine youth tobacco access laws on the rate of change in smoking prevalence among high school-aged youth.**

***Index Terms*—linked data, smoking, public health**

I. Introduction

With the unprecedented ability to store, share, and utilize data afforded by the web, coupled with the federal government's digital government strategy encouraging broad access to government-funded data sources,[1] opportunities to explore, compare, and utilize all kinds of data abound. We previously sought to improve access to, utilization, and understanding of population science data, that could shed light on public health intervention strategies and health behavior patterns. In order to do this effectively, a number of data sets were integrated by making their interrelationships more explicit for data harmonization. One goal of this effort was to enable new analyses of complex behaviors and systems [1,2] to assist researchers and policymakers in exploring hypotheses on correlations between intervention strategies and health outcomes.

The current project aims to provide access to publicly available datasets including the ImpacTeen Tobacco Control Policy and Prevalence Data: 1991-2008,[2] the Community Health Status Indicators (CHSI) survey[3], and the Heath Information National Trends Survey (HINTS) [3]. This has the potential to enable a broad range of users to explore and understand data that can possibly drive decision making by both governments and individuals. In the broader biomedical community, conventional human-readable data dictionaries are moving online. These online dictionaries serve as controlled vocabularies – providing the potential for term usage to become better understood. As vocabularies move across the ontology spectrum [4] to include more detailed computer understandable descriptions of meaning, computer supported tools for "understanding" relationships between datasets become more possible. Recognition of the need for standardized measures, controlled vocabularies, and data dictionaries in the population sciences is increasing, with existing efforts

[1] http://www.data.gov/about

[2] http://impacteen.org/tobaccodata.htm

[3] http://www.communityhealth.hhs.gov/

covering a rich set of descriptions across multiple scientific disciplines [5,6]. Recent efforts have been made [7] that create information models that are annotated with semantic metadata to be added to the dataset. This has been accomplished with a subset of the HINTS dataset [8], but is an expensive process in both time and expertise. More detailed descriptions of term meanings are especially useful as applications begin to draw inferences across data sets. A number of health and life science ontologies have reached a level where large communities have agreed upon term meanings and the vocabularies and meanings are stable, reasonably well understood, and the ontologies are reasonably stable. Collections of many well-used ontologies exist in places such as the NCBO Bioportal [9], the Open Biological and Biomedical Ontologies collection at OBO Foundry [10], and other places. Some of these vocabularies and ontologies can provide a reference set of knowledge structures to which the terms in datasets may be referenced. However for many scientists, bridging datasets using even the best and most easily understood of these representations remains a non-trivial task.

Additionally, while many data producing efforts seek to provide good metadata describing the datasets the generation of metadata can be time consuming and often incentives for good production are lacking. We hope to provide one incentive for good metadata production by showing that improved metadata, through a Semantic Data Dictionary, results in improved capabilities using tools like qb.js. This in turn lead to greater insights about the data and the ability to better communicate those insights to the rest of the world.

II. Background

Previously, we proposed the PopSciGrid health portal, and have developed some initial views of population data related to tobacco behavior and policy. In previous papers [8,13], we described some of the benefits in using semantic applications to enable data integration, and in particular we described how this was accomplished in the domain of population science related to tobacco policies and smoking prevalence. We began to incorporate Linked Data principles and tools and included smoking prevalence datasets from impacteen.org [8] as well as smoke-free policies from the National Cancer Institute to be used for the conversion, integration, provenance, representation, and visualization of data. We used conversion tools from RPI [11] and encoded provenance initially using the Proof Markup Language (PML) [12]. Today's encodings utilize the emerging World Wide Web Consortium (W3C) proposed recommendation for provenance – PROV and PML.

That effort was a success in that the broad and diverse team successfully integrated and visualized relevant tobacco policy, demographic, and behavioral data and maintained provenance. However, we wanted users to investigate and develop hypotheses based on specified data sets without requiring knowledge of linked data or semantic tools. We investigate its use in analyzing tobacco use interventions, taxation, and smoking bans. This will enable users to explore data that can inform them as they make decisions about behaviors that may impact their health. The team also designed a consumer health portal that would allow the public to explore how states or regions may compare to others as well as the national trends in risk factors for cancer and related behaviors such as tobacco use, exercise, balanced diet, and cancer screenings [13]. This can be advantageous for a non-specialist in creating and exploring visualized data.

While the original focus of tobacco prevalence and tobacco policies has remained consistent, we needed to create tools that minimized requirements for users who lacked sophisticated expertise in semantic technology. As a result, our team has since developed qb.js, a tool that permits users to investigate and develop hypotheses based on specified data sets without requiring knowledge of linked data or semantic tools. We investigate its use in analyzing tobacco use interventions, taxation, and smoking bans. This will enable users to explore data that can inform them as they make decisions about behaviors that may impact their health. The team also designed a consumer health portal that would allow the public to explore how states or regions may compare to others as well as the national trends in risk factors for cancer and related behaviors such as tobacco use, exercise, balanced diet, and cancer screenings [13]. This can be advantageous for a non-specialist in creating and exploring visualized data. We have expanded our research by incorporating more detailed visualizations using semantic technology to increase the interactivity of health data at the disposal of the public. Our current focus is examining the connections between health, behavior, policy, and demographic data, while analyzing the relationship between policies over time that ban the use of tobacco in public spaces and the number of consumers that may or may not be impacted by the enforcement of such policies. The PEW Internet & American Life Project's survey results revealed that the public is constantly seeking sources in order to monitor government activities, assess the impacts of new legislation, and to observe the allocation of their tax dollars [11]. This is significant to our research, as it shows an obvious interest by the public in government affairs. It also demonstrates why semantic technologies are useful in disclosing the information about such affairs.

We argue that a linked open government data (LOGD) approach can be useful for helping to connect related content and also reducing expenses that are incurred from distributing health data on non-web oriented media. The integration of semantic web and linked data principles in government data can increase the transparency in the relationship between the public and government. This in turn grants the public greater autonomy over health issues and allows public health officials to display health information in a more structured and organized manner. With this sort of infrastructure in place, programs, and initiatives such as Data.gov can be more efficient by allowing the public to be collectively involved with government data access from different networks, crafting applications, and providing constructive criticism to improve the quality of the distributed government data.

To validate the use of these semantic technologies, we embarked on an analysis of a widely used dataset from Impacteen.org called “Tobacco Control Policy and Prevalence Data: 1991-2008” and began by loading this data into qb.js[4] (see Figure 1) and into a motion chart visualization[5] (see Figure 2). We immediately noticed that there seemed to be a decrease in the prevalence of smoking among youth after implementation of some youth tobacco access laws, but could not tell if it was because of ongoing decreases in tobacco use prevalence. We therefore determined that we needed to perform a statistical test to see if changes in tobacco access laws reliably were correlated with subsequent changes in the rate of change in youth smoking.

III. Related Work

Paulheim et al. show how to use Linked Data resources to generate hypotheses that explain particular statistics [14]. Riedewald et al., proposed a modular framework that combined one-dimensional On-Line Analytical Processing (OLAP) aggregation methods in order to better enhance the data cube’s ability to reserve space for information [15]. Two dimensional OLAP was proposed in 2011 by Ordonez et al., that displays interactive visualized data which separates and distinguishes the differences in statistical measurements [16]. This critical distinction is one that we leverage in qb.js. The RDF Data Cube Vocabulary was designed to explicitly support OLAP-like operations on RDF data, and qb.js is designed to work with the RDF Data Cube Vocabulary.

We build on a long line of work in formalizing data dictionaries. Cimino emphasized the importance of the knowledge based terminology in data dictionaries in order to make it more convenient for physicians to understand and apply their knowledge to the way in which they care for their patients [17]. Hinds introduced the semantic data dictionary for the NASA, which has established the foundation for further semantic technological advancements [18]. For example, a medical researcher might ask: how does air-quality effect emphysema? This and many similar questions will require sophisticated semantic data integration. The researcher who raised the question may be familiar with medical data sets containing emphysema occurrences. But this same investigator may know little, if anything, about the existence or location of air-quality data. It is easy to envision a system which would allow that investigator to locate and perform a ‘‘join’’ on two data sets, one containing emphysema cases and the other containing air-quality levels. No such system exists today. Similarly, Ruan et al. explain the convenience of semantic networks because of the ability to bridge medical terminology and information sources [19]. Herrea et al., however, have an alternative approach that encourages the use of ontologies for the purpose of strengthening the semantic groundwork of distributing health data [20]. These results highlight the fact that structure and organization are still factors in crafting authentic computer-based software and applications that will efficiently present health information. Much of this research also depicts why it is essential for the general public to have a greater in-depth understanding rather than there being a veil between them and policy-makers.

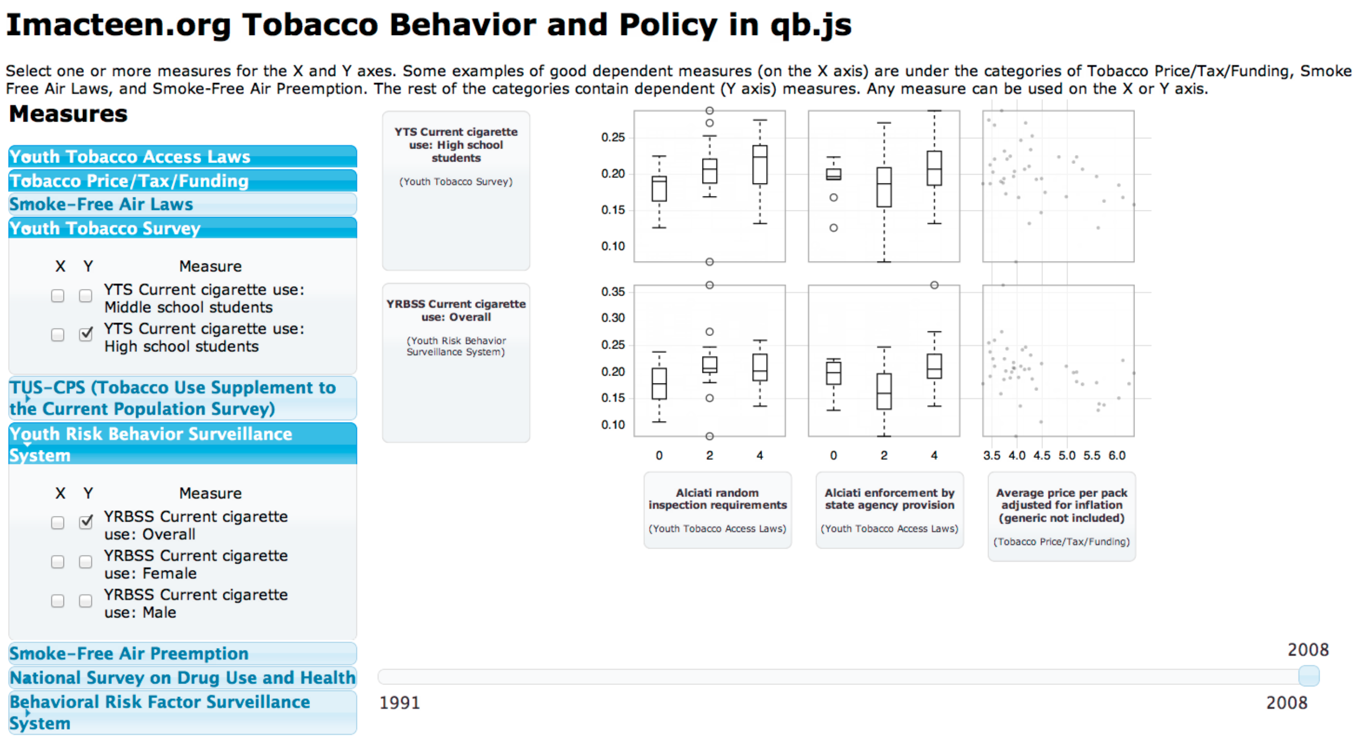
While much of our research is comprised of advocating for the use of semantic technologies to display health data, we are specifically analyzing how tobacco access policies impact the prevalence of smoking by using semantic technology to explore the data and formulate an analysis. Tworek et al., performed a population-based study, showing that tax increases on tobacco products is an effective method in tobacco control policy by decreasing smoking prevalence and increasing smoking cessation among youth [21]. The numbers of smoker reduce because of the enforced tobacco policies that discourage cigarette consumption. Alciati et al., designed a rating system that evaluated and compared the states that strictly enforced policies to prevent adolescent youth from gaining access to tobacco products with states that passively enforced those policies. Alciati’s results revealed that state legislatures need to institute more effective and relevant regulations that limit the alternative avenues that enable youth access to tobacco [22]. Wakefield et al., provides supporting evidence with the results of a national survey revealed that when there are firm restrictions against cigarette smoking the prevalence of smoking among teenagers decreased substantially [23]. Similarly, Botello-Harbaum et al. found that adolescent students in states with strict regulations in comparison to adolescent students that lived in states with little to no restrictions were not as like to be regular smokers, and concludes that prices are huge factors in deterring youth from purchasing and consuming tobacco products [24]. These results prove that enforced policies and convenient access to health data can impact the health decisions of the public.

IV. Methods

We sought to validate the use of the RDF Data Cube Vocabulary and qb.js for initial exploration of data and for performing statistical analyses based on the questions we asked based on that data. Towards that end, we developed a statistical method for comparing the effects of policy changes at the state level to take advantage of behavioral and policy measures published in the ImpacTeen Smoking dataset. Since most of the measures of interest are available longitudinally, it was clear that the data could be split for each state before a policy was implemented and after it was implemented. A linear regression model is created for each state before and after the change in policy in the behavioral measure in question for each policy under analysis, for each state. We use the major coefficient from the linear model to represent the rate of change in tobacco behavior in the population in question. We examine rates of change because long-term trends in tobacco use have been shifting over time and we would like to capture any additional transformations that occur after a policy change. We then use a state-by-state paired t-test to determine if there are significant differences in behavior after a given policy is

[4] http://orion.tw.rpi.edu/~jimmccusker/qb.js/examples/tobacco

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



Imacteen.org Tobacco Behavior and Policy in qb.js

Select one or more measures for the X and Y axes. Some examples of good dependent measures (on the X axis) are under the categories of Tobacco Price/Tax/Funding, Smoke-Free Air Laws, and Smoke-Free Air Preemption. The rest of the categories contain dependent (Y axis) measures. Any measure can be used on the X or Y axis.

Measures

Youth Tobacco Access Laws

Tobacco Price/Tax/Funding

Smoke-Free Air Laws

Youth Tobacco Survey

X	Y	Measure
□	□	YTS Current cigarette use: Middle school students
□	☑	YTS Current cigarette use: High school students

TUS-CPS (Tobacco Use Supplement to the Current Population Survey)

Youth Risk Behavior Surveillance System

X	Y	Measure
□	☑	YRBSS Current cigarette use: Overall
□	□	YRBSS Current cigarette use: Female
□	□	YRBSS Current cigarette use: Male

Smoke-Free Air Preemption

National Survey on Drug Use and Health

Behavioral Risk Factor Surveillance System

YTS Current cigarette use: High school students
(Youth Tobacco Survey)
YRBSS Current cigarette use: Overall
(Youth Risk Behavior Surveillance System)
0.25
0.20
0.15
0.10
0.35
0.30
0.25
0.20
0.15
0.10
0
2
4
0
2
4
3.5
4.0
4.5
5.0
5.5
6.0
Alciati random inspection requirements
(Youth Tobacco Access Laws)
Alciati enforcement by state agency provision
(Youth Tobacco Access Laws)
Average price per pack adjusted for inflation (generic not included)
(Tobacco Price/Tax/Funding)
2008
1991
2008

Figure 1. Two tobacco control policies and two tobacco behavior measures are displayed using qb.js. This is displaying data for 2003. It is unclear if the changes in the bar graph are the result of a response to policy change or are the reason the policy was needed. The Alciati measures are a ratings scale where higher numbers represent increased restrictions on youth tobacco access for a particular strategy. Average price per pack is in US dollars, and cigarette use values are expressed as a fraction of the general population. http://orion.tw.rpi.edu/~jimmccusker/qb.js/examples/tobacco/

implemented. Only policies that have been changed in at least two states are analyzed.

The paired t-test gives us a unique level of control over experimental variables. By pairing the states before and after the change, we effectively record the experiment that each state legislature is implicitly performing on their state. Through this, we gain control for geographic, demographic, socioeconomic, and cultural biases due to the relative stability of culture and socioeconomics that states have during the time scales under study. Any long term shifts in a given state that happen to co-occur with the change in policy that could result in changes in smoking prevalence would require co-occurrent changes in most of the studied states to have an impact in the statistics, and if it were due to chance it would simply decrease the statistical power of the test without interfering with the overall results.

We developed an R script[6] that queries the LOGD SPARQL endpoint for the data using the measures in question. It then creates a derived dataset that shows longitudinal differences between changes in policy for each state and policy. The software filters out the datasets without at least two years of data before and after changing the law measurement to allow for a valid computation of rate of change. This is acceptable because the sample size for each year is significant enough to assure the reliability of that data within a given year. The coefficients before and after the policy change are then compared in a paired t-test for each state that has had a change in policy during the period that the behavioral measure has been recorded. The derived dataset is saved as a Comma-Separated Value (CSV) file and is converted to Resource Description Framework (RDF) format by the R script. Both the resulting queries and analyses are straightforward and are assisted by the RDF Data Cube Vocabulary.

All data and measure metadata were encoded using the RDF Data Cube vocabulary[7] to allow for integration into qb.js and to facilitate exploration and analysis of the data. All data has been loaded into the RPI Tetherless World Constellation's LOGD SPARQL endpoint at http://logd.tw.rpi.edu/sparql. All URIs referenced in the rest of this paper have information encoded in this database and are accessible in the graph http://logd.tw.rpi.edu/source/impacteen-org/dataset/tobacco-control-policy-and-prevalence/version/2012-Jan-16. For a behavioral measure of youth tobacco use we identified the Youth Risk Behavior Surveillance System (YRBSS) as a good indicator for smoking among high school-aged students because it had the longest available records of prevalence data for each state. We use the Overall measure,[8] which is the "overall percentage of youth who reported past month

[6] http://bit.ly/MDYR7c

[7] http://w3.org/TR/vocab-data-cube

[8] This measure is encoded by the URI http://health.tw.rpi.edu/source/impacteen-org/dataset/tobacco-control-policy-and-prevalence/measure/YRBSS_Current_cigarette_use_Overall

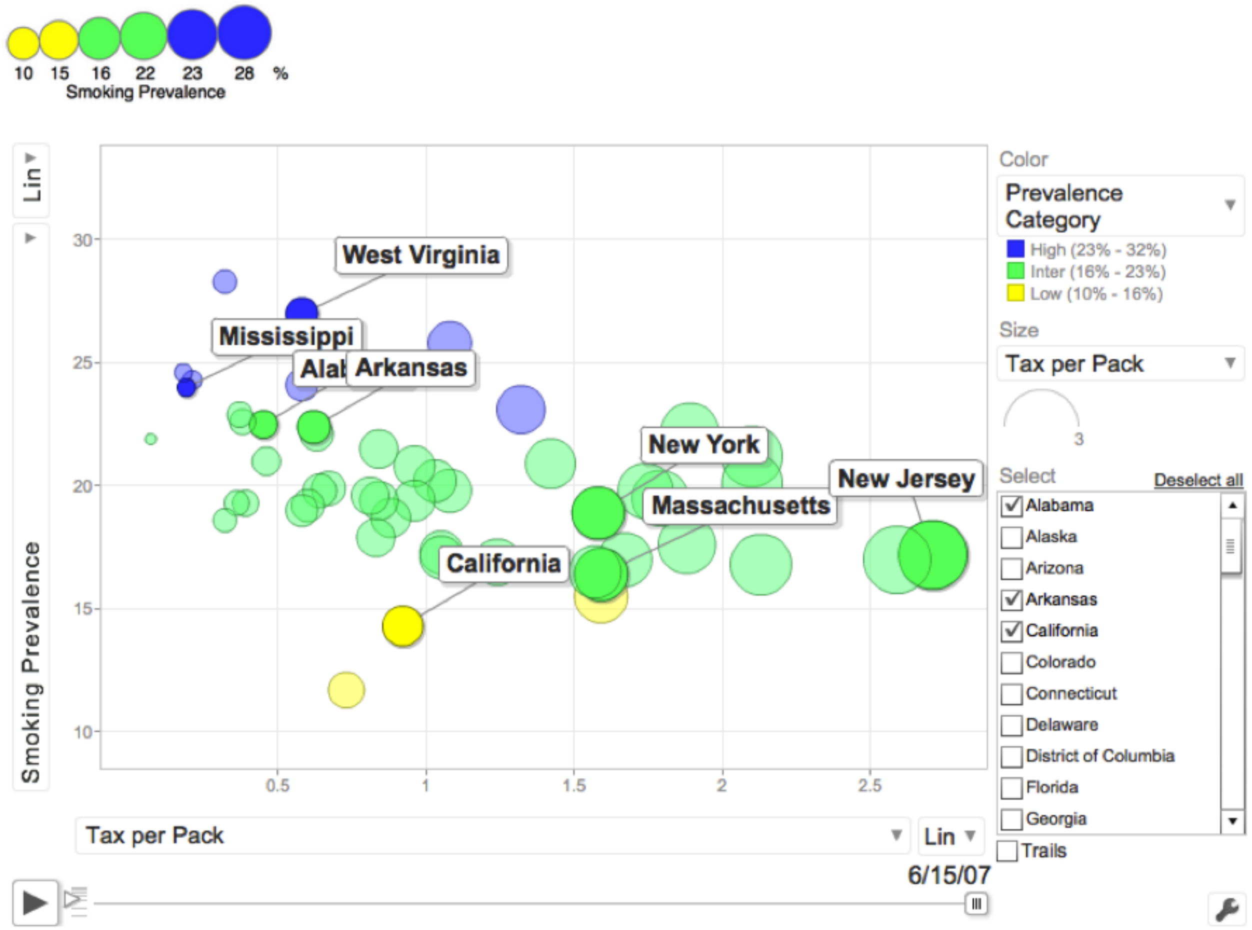


Figure 2. A motion chart showing the relationship between Tax per Pack and Smoking Prevalence. As taxes go up, the prevalence goes down, but it is unclear if this is because smoking prevalence is going down in the population in general or only in youth.

cigarette use in the CDC‘s YRBSS survey conducted among students in grades 9-12 in selected states. Samples are representative of jurisdictions and include predominantly public schools, though private schools are included for some samples."[9] Other measures, such as the National Survey on Drug Use and Health (NSDUH) and the Youth Tobacco Survey (YTS) cover much shorter longitudinal spans, which would result in fewer states that could be analyzed for changes in the rate of change in smoking behavior. Other surveys did not offer data on youth tobacco behavior.

The policy measures were simple to identify, as the Alciati measures for youth tobacco access laws [12] as well as the Possession-Use-Purchase measures [23] were encoded in the ImpacTeen dataset.[10]

V. Results

This analysis was performed on 11 policies that met the above criteria. As reported in Table 1, 9 of those policies had a statistically significant effect on the rate of change in smoking among high school-aged students. The remaining two may suffer from a lack of data, as their mean effect is in line with the other policies measured. These are consistent drops of on average 2-3% per year in smoking rates among high school-aged students when a state implements that particular policy. This is not simply a one-time drop in prevalence, but is instead a consistent and significant drop in prevalence for every year measured after the change.

Additionally, the hypotheses raised in this paper were a direct result of explorations and questions raised through exploration that was conducted using qb.js. We had observed that there seemed to be a relationship between lower smoking prevalence in High School Students and certain youth tobacco access policies, but the visual analysis made it clear that a better longitudinal analysis was needed to settle the question. We were then able to formulate our hypotheses and analysis very easily based on exploration using qb.js.

[9] From the ImpacTeen Tobacco codebook: http://impacteen.org/statetobaccodata/Codebook.pdf

[10] These measures are all listed as parts of (http://www.obofoundry.org/ro/ro.owl#part_of) the Youth Tobacco Access Laws measure category (http://health.tw.rpi.edu/source/impacteen-org/dataset/tobacco-control-policy-and-prevalence/category/Youth_Tobacco_Access_Laws)

Policy	*# of States*	*Mean change in prevalence rate change*	*p-value*
Alciati random inspection requirements	8	-3.11%	0.0013
Alciati enforcement by state agency provision	8	-2.44%	0.0027
Alciati restrictions on packaging	10	-2.05%	0.0075
PUP Minors purchase prohibited	6	-2.39%	0.0207
PUP Minors use prohibited	5	-3.41%	0.0239
PUP Minors possession prohibited	6	-2.88%	0.0265
Alciati free distribution restrictions	5	-2.85%	0.0309
Alciati clerk intervention requirement	4	-2.29%	0.0357
Alciati vending machine restrictions	5	-2.18%	0.0387
Possession-Use-Purchase Index	5	-2.29%	0.0581
Alciati penalties to retailers	3	-3.60%	0.0729

Table I
Effects of Tobacco Policies on the rate of change in High School student smoking prevalence. Of the analyzed policies, only Possession-Use-Purchase Index, a composite of other policies, and Alciati penalties to retailers do not show a significant change. However, this may have more to do with the available sample size of states than the invalidity of the policies.

VI. Discussion

This innovative methodology allows exploration of existing datasets in a powerful way. It illustrates, for example, a clear 2-3% effective boost to the decline of smoking in high school students when these policies are put into effect. In light of these results, public health policymakers should consider greater exploration of the potential impact of enacting such policies on high school student smoking rates. These results would have been much more difficult to obtain if the authors had used conventional analysis tools and data representation techniques. The RDF graph structure, even without semantics in place, makes it less complex to extract relevant slices of datasets for analysis. The existence of tooling for statistical languages like R to query large RDF graphs using the SPARQL language will free up analysts to focus on the data and development of better analytical methods.

We also seek to encourage producers of data to use the RDF Data Cube representation of scientific data because it provides a means to create a Semantic Data Dictionary based on the RDF Data Cube Vocabulary and other ontologies such as the Measurement Units Ontology and the World Wide Web Consortium's Provenance Ontology (PROV-O). Other tools such as the RDF Data Cube and the R scripts used here are provided as potential incentives for creating and providing such representations. Use of qb.js should only be limited to the ability to represent data in RDF and to provide metadata using semantic data dictionaries. Semantic data dictionaries are easily composed from the prose data dictionaries that generally accompany published datasets. When a conventional data dictionary is missing, it becomes difficult for any user of the data to interpret it, whether they use semantic technologies or not.

VII. Future Work

Although we have argued that qb.js is an efficient tool that has potential to enhance the display of statistical health data, there is further work to be done. For instance, supporting existing efforts to establish standardized measures and controlled vocabularies [1,2], and establishing a data dictionary editor would be useful for revising the definitions that are provided for the medical terminologies. Even before establishing a semantic data dictionary editor, there are questions that remain, such as how best to convert conventional data dictionaries into annotated semantic data dictionaries using RDFa. In addition to converting annotated semantic dictionaries, a tool must be designed a manner that it is accessible to communities that are both specialized and non-specialized in the field of semantics.

Furthermore, we hope to take advantage of CKAN for potential publication and production of these datasets, semantic data dictionaries, and visualizations. We intend for CKAN publications to make it much simpler for non-specialists to create the needed data representations so that users can more easily take advantage of tools like the RDF Data Cube. By doing so, our goal of increasing transparency and accessibility of health data between the public and policy-makers will be achieved.

VIII. Conclusions

We have introduced the data cube as method of allowing a wide range of users to access and explore statistical health data. We have provided some examples of the benefits of including hypothesis formation followed by relatively easy exploration of and visualization of the hypothesis. We have demonstrated the benefits of using qb.js through our analysis of the effect of youth tobacco access laws on high school smoking prevalence. The ability to easily access and manipulate data is the first step in helping the public understand how their lives maybe impacted by health-related policies and regulations.

Finally, the use of semantic technologies in projects like LOGD provides a pathway for scientists to better describe their data in ways that help the visualization, exploration, and analysis of data by the data producers as well as consumers. We encourage the use of qb.js and the development of similar tools to increase the utility of well-described data to scientists, policy makers, and the public at large.

References

[1] R. Moser, B. Hesse, A. Shaikh, P. Courtney, G. Morgan, E. Augustson, S. Kobrin, K. Levin, C. Helba, D. Garner, et al. "Grid-Enabled Measures: Using

Science 2.0 to Standardize Measures and Share Data" American Journal of Preventive Medicine vol. 40:5 pp S134-S143 (2011).

[2] P. Stover, W. Harlan, J. Hammond, T. Hendershot, and C. Hamilton, C.M. "PhenX: a toolkit for interdisciplinary genetics research" Current opinion in lipidology vol. 21:2 p 136 (2010).

[3] B. Hesse and R. Moser. Health Information National Trends Survey (HINTS), ICPSR25262-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], (2007.) doi:2009-06-23. doi:10.3886/ICPSR25262.v1

[4] P. Schad, L. Mobley, and C. Hamilton. "Building a Biomedical Cyberinfrastructure for Collaborative Research" American journal of preventive medicine vol. 40:5 pp. S144-S150 (2011). doi:10.1016/j.amepre.2011.01.018

[5] R. Moser, B. Hesse, A. Shaikh, P. Courtney, G. Morgan, E. Augustson, S. Kobrin, K. Levin, C. Helba, D. Garner, M. Dunn, and K. Coa "Grid-Enabled Measures: Using Science 2.0 to Standardize Measures and Share Data" American journal of preventive medicine vol. 40:5 pp. S134-S143 (2011) doi:10.1016/j.amepre.2011.01.004

[6] D.L. McGuinness "Ontologies come of age" Spinning the semantic web: bringing the World Wide Web to its full potential The MIT Press p 171 (2005).

[7] K. Buetow "Cyberinfrastructure: empowering a" third way" in biomedical research" Science, 308:5723 pp 821-824 (2005).

[8] D.L. McGuinness, T. Lebo, A. R. Shaikh, R. P. Moser, L. Ding, J. P. McCusker, G.D. Morgan, Z. Tatalovich, G. Willis, and B.W. Hesse, "Towards Semantically-Enabled Next Generation Community Health Information Portals: The PopSciGrid Pilot", 45th Hawaii International Conference on System Sciences, 2752-2760, 2012.

[9] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. B. Griffith, C. Jonquet, D. L. Rubin, B. Smith, M. A. Storey, C. G. Chute, M. A. Musen, "Bioportal: Ontologies and Integrated Data Resources at the Click of a Mouse". Nucleic Acids Res, pp. W170-173, 2009.

[10] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, The OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration", Nat. Biotechnol, (2007), pp. 1251–1255

[11] L. Ding, T. Lebo, J.S. Erickson, D. DiFranzo, G.T. Williams, X. Li, J. Michaelis, A. Graves, J.G. Zheng, Z. Shangguan, J. Flores, D.L. McGuinness, J. Hendler, TWC LOGD: A Portal for Linked Open Government Data Ecosystems, Web Semantics: Science, Services and Agents on the World Wide Web , 2011, p.p. 1-12.

[12] D. McGuinness, L. Ding, P. Pinheiro Da Silva, and C.Chang, "PML 2: A modular explanation interlingua," in Proceedings of AAAI, vol. 7, (2007).

[13] 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 Proceeding of the Third International Web Science Conference, Koblenz, Germany, June 15-17 2011, p.p. 1-3.

[14] H. Paulheim, E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti. "Generating Possible Interpretations for Statistics from Linked Open Data" Lecture Notes in Computer Science, (2012) vol. 7295, pp 568-574 doi:10.1007/978-3-642-30284-8_44

[15] M. Riedewald, D. Agrawal, and A. E. Abbadi, "Flexible Data Cubes for Online Aggregation,", ICDT, 2001, pp. 159-173

[16] C. Ordonez, Z. Chen, and J. García-García, "Interactive Exploration and Visualization of OLAP Cubes", Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP, 2011, G

[17] J. J. Cimino, "From Data to Knowledge through Concept-oriented Terminologies: Experience with the Medical Entities Dictionary", J Am Med Inform Assoc. 2000, pp.288–97.

[18] N. Hinds, Y. Huang, and C. V. Ravishankar. "A semantic data dictionary method for database schema integration in CIESIN", In Proceedings of the Conference on Earth and Space Science Information Systems, 1992, pp. 139-151.

[19] W. Ruan, T. Bürkle, J. Dudeck, "An Object-Oriented Design for Automated Navigation of Semantic Networks Inside a Medical Data Dictionary", Artif Intell Med, 2000, pp.83-103.

[20] H. Herrea, B. Hellera, "Semantic Foundations of Medical Information Systems Based on Top-Level Ontologies", Knowledge Based Systems, 2006, pp.107–115.

[21] C. Tworek, R. Yamaguchi, D. D. Kloska, S. Emery, D. C.,Barker G. A. Giovino, P. M. O'Mally, and F. J. Chaloupka, "State-Level Tobacco Control Policies and Youth Smoking Cessation Measures", Health Policy, 2010, pp.136–44.

[22] M. H. Alciati, M. Frosh, S. B. Green, R. C. Brownson, P. H. Fisher, R. Hobart, A. Roman, R. C. Sciandra, and D. M. Shelton, "State Laws on Youth Access to Tobacco in the United States: Measuring their Extensiveness with a New Rating System", Tobacco Control, 1998, pp.345–352 doi: 10.1136/tc.7.4.345

[23] M. A. Wakefield, F. J. Chaloupka, N. J. Kaufman,

C. T. Orleans, D. C. Barker, E. E.Ruel, “Effect of Restrictions on Smoking at Home, at School, and in Public Places on Teenage Smoking: Cross Sectional Study” BMJ, 2000, pp.321:333

[24] M.T. Botello-Harbaum, D. L. Haynie, R. J. Iannotti, J. Wang, L. Gase, and B. Simons-Morton, “Tobacco Control Policy and Adolescent Cigarette Smoking Status in the United States", Nicotine Tobacco Research, (2009) pp. 875-885