A Semantic Web Representation of Entire Populations

Daniel Welch, Amanda Hicks, Josh Hanna, William R. Hogan
Department of Health Outcomes and Policy
University of Florida
Gainesville, FL
dwelch2101@ufl.edu, aehicks@ufl.edu, joshhanna@ufl.edu, hoganwr@ufl.edu

Abstract—Accurately representing demographic realities is a critical component in creating useful, agent-based epidemiological models of infectious disease. Synthetic ecosystems are generated from Census data microsamples in a statistically-sound manner to maintain population-level demographic characteristics. These highly detailed representations of populations are the basis of many advanced simulations of infectious disease epidemics. Creating a standard, machine-readable representation of synthetic ecosystem data would enable easier use and integration with epidemic simulator software. Here we describe an ontology-based representation in Resource Description Framework (RDF) and Web Ontology Language (OWL) of version 1.0 of the 2010 U.S. Synthetic Population database by RTI International. Our representation draws upon applicable classes from several reference ontologies, including the Ontology of Medically Related Social Entities (OMRSE). After failing to find suitable ontological representations of several key data elements in the Synthetic Population dataset, we created new classes in OMRSE for representing employment status, employee roles, workplaces, residences, households, and age measurements. We loaded a test RDF dataset (structured according to ontologies in OWL) of synthetic individuals into a commercial triple store (Stardog) and validated the representation with SPARQL queries.

Keywords—ontology; synthetic ecosystem; disease transmission model

I. BACKGROUND

Disease transmission models (DTMs) are epidemiological models that predict the future course of infectious disease outbreaks under various assumptions. They are used to study which strategies for controlling outbreaks are potentially the most effective and thus for decision making during the course of an outbreak. For example, researchers have used them to study the effects of various vaccination control strategies on pandemic influenza [1, 2, 3] and to study the Ebola outbreaks in western Africa [4].

Agent-based disease transmission models (AB-DTMs) are a class of DTMs that represent every host individual in the population of interest, and sometimes individual vector organisms as well, to increase the realism of the simulations and thereby increase the accuracy of the predictions generated [5]. To accomplish these goals, the characteristics of the population represented in the AB-DTM must closely match the characteristics of the actual population under study. As reported by Grefenstette et al., accounting for differences in population density and sociodemographics indeed affects DTM results for different regions [5].

Census data are a key resource for matching simulated populations to actual ones [5, 6]. They include data about demographics, housing units, household composition, employment, school attendance, and other physical and social dynamics that have the potential to influence infectious disease transmission. However, record-level Census data are typically only available as microsamples of the overall Census data set. Therefore, there are typically representations of only 1%-5% of the population. This amount of data is insufficient for use with AB-DTMs that model 100% of a population. To overcome this limitation, researchers employ statistical methods to generate a full population dataset from the microsamples such that the synthetic population-wide dataset mirrors the actual population in aggregate in terms of various demographic characteristics such as sex, race, and marital status [6]. For example, the synthetic populations (or more generally, synthetic ecosystems, since housing units, workplaces, schools, etc. are also represented) available have percentages of blacks, women, employed individuals, students, etc. that statistically match the actual population.

In addition to expanding microsample Census data to full population size, researchers incorporate significant additional information into synthetic ecosystems relevant to disease transmission [5, 6, 7]. For example, although Census data capture employment and school attendance statuses, they do not associate individual persons to individual workplaces or schools. However, for AB-DTMs, these linkages are critical for studying whether and how well school closures and workers' decisions to stay at home (whether made individually or as a matter of public health or employer policy) control disease transmission. Therefore, a significant component of extant synthetic ecosystems is data about individual school and workplace assignment.

Researchers typically make these synthetic ecosystems available as delimited text files in a format suitable for loading into tables in a relational database management system, with limited semantics and simple integer values for representing categories such as race and gender. For example, see the extensive collection of synthetic ecosystems available at [8].

In this work, we make available full-population synthetic ecosystem data as Resource Description Framework (RDF) [9] triples. It differs from past efforts to include Census data in government linked open data (LOD) [10] in at least two key respects. First, we took a realist ontological perspective. To our knowledge, our work is the first to attempt to represent the necessary entities to cover a Census-derived dataset from a realist perspective. We were able to reuse significant components of other realist-based ontologies, but we also needed to carry out additional ontology development to accomplish the task. Second, to our knowledge, we are the first to attempt representing an entire population from Census data in a Semantic Web framework using synthetic ecosystem data created for AB-DTMs.

In previous work, we created the Ontology of Medically Related Social Entities (OMRSE) to handle demographics such as those represented in Census and electronic health record (EHR) data [11]. OMRSE is a realist representation of medically related social entities. Social entities are those entities that exist in reality but which would not exist outside of a social context. For example, the role of a doctor is distinct from the human being who bears that role. This role exists within the healthcare system and confers rights and responsibilities associated with treating and diagnosing a patient. It is the result of social agreements and interactions rather than of the physical stuff that makes up the natural world. It is realized through various processes of diagnosing, treating, prescribing, etc. We develop OMRSE in accordance with OBO Foundry best practices [12] and reuse classes from several other ontologies including Basic Formal Ontology [13], NCBI Taxonomy [14], Information Artifact Ontology [15], and the Document Acts Ontology [16].

Given the importance of school and workplace assignment and the data about them in synthetic ecosystems, it was critical to represent additionally the roles of students and employees. Furthermore, it was necessary to capture the relationships of these roles to the organizations that create them and to the individual facilities where they are realized. We also report here on the extent to which pre-existing ontologies fulfilled this need vs. the additional ontology development required.

II. Methods

We reviewed the files generated by the Research Triangle Institute's Synthia synthetic population generator [6] in conjunction with its documentation. Because Synthia uses U.S. Census files and public-use microsample (PUMS) data, we also reviewed U.S. Census definitions of the variables in those data.

We reviewed each of the data fields in the following subset of Synthia files: *synth_people.txt*, *synth_households.txt*, *schools.txt*, *workplaces.txt*. Through an iterative process, we analyzed and described each data field and determined whether to include the field in this work. The most common reason we excluded a field from the final ontological representation was redundancy. For example, we excluded data fields from *synth_households.txt*, *schools.txt*, and *workplaces.txt* that represented the total number of individuals assigned to a household, school, or workplace since these values could be derived by counting in the underlying data. Other fields were excluded because they were determined to be of lesser immediate importance to epidemiologic simulation, such as the *prek, kinder, gr01-12,* and *ungraded* fields in *schools.txt*, which represent the total number of students in different grade categories in a given school.

We then determined whether each included data field could be accurately modeled using existing ontological classes from OMRSE or other established ontologies, or whether new classes were necessary. We created graphical models of how the data would be structured ontologically as an initial specification for transforming the data into RDF, as well as to identify any new classes that we would need to create. These graphical models depict the individuals, relationships between pairs of individuals, and classes to which individuals belong. These diagrams included specifications for associating people with their workplaces and schools as represented in the dataset.

We then manually created these individuals and relationships for a single set of individuals in one household, including their associated school and workplace, in a Web Ontology Language (OWL) [17] file that imported OMRSE and the Apollo-SV ontology [18] (the latter was a choice of convenience because it already brings together ontological representations from numerous ontologies, including its own, in the domain of epidemic simulation). This OWL file served as the machine-readable specification for converting Synthia text files into RDF triples. Once we had this machine-readable specification, we created a software application that performed this conversion, and applied it to the county-based Synthia files for Alachua County, FL and Miami-Dade County, FL. This application is freely available at: https://github.com/ufbmi/synthia-rdf-converter. We then loaded the RDF triple datasets output by the application into an instance of the Stardog triple store.

A. New Ontology Classes in OMRSE

In accordance with OBO Foundry best practices, we reused as many classes and object properties from other ontologies as we could to generate the OWL file. After importing existing classes from OBO ontologies, it was still necessary to create new classes to represent several key elements of the Synthetic Population dataset. Specifically, we created new classes in OMRSE to represent employment status, employee roles, workplaces, residences, households, and age measurements.

B. Queries of the RDF Dataset

We developed queries of the RDF datasets to validate our representations as well as to identify population characteristics that are likely to influence disease transmission. If these differences are signficant among regions, they could influence the choice of DTM used to study an infectious disease control strategy. For example, if two regions differ substantially in household and workplace composition, size of school-aged population, etc., an AB-DTM is likely to be the better choice. Furthemore, these queries could also be done as part of a simulation experiment to help explain differing results among geographical regions in incidence rates, peak dates, and choice of infectious disease control strategies output by the simulator.

Because the sizes of households, schools, workplaces, and the amount of overlap among them (e.g., households with an employee in the workplace and student in a school) influence disease transmission and thus potentially DTM results, we

developed queries to find (1) the average numbers of individuals per household, workplace, and school; (2) the number and percentage of households with both an employee and a student; and (3) the number and percentage of workplaces with at least one employee who lives with a student. We executed these queries against both the Alachua County and Miami-Dade County datasets to contrast these locations based on characteristics relevant to disease transmission.

We loaded the RDF data into an instance of version 3 of the Stardog triple store from Complexible, Inc. This triple store runs on an Amazon Web Services r4.large instance (2 CPUs and 15.25GB of RAM). Queries were submitted from the Stardog command line on the same server on which the triple store was running. The timings we report here are from the Stardog command line output.

III. Results

A. RDF Representations

To accurately model the U.S. Synthetic Population Database, we created RDF representations of the data fields relating to individual persons, households, housing units, workplaces, and schools. We created graphical models of these representations (Figs. 1-4). Fig. 2 illustrates our representation of humans in a household. Fig. 3 illustrates our representations of workplaces and employment.

Many ontologies classify age as a physical quality, rather than as a measurement of some temporal interval with respect to the time the measurement was made. The Ontology for Biomedical Investigations (OBI) [19] has a class 'age measurement datum' that has a class restriction of being *is about* some age quality. The age quality class, in turn, comes from the Phenotypic Quality Ontology (PATO). By contrast, we represent age as a measurement of a one-dimensional temporal region that is occupied by a process that is part of the history of some object (Fig. 4).

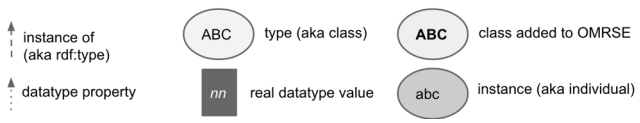

Fig 1. Key for Graphical Models.

In analyzing Synthia data fields, we found that Synthia conflates households and housing units, despite being based on U.S. Census data that make the distinction clear. For example, Synthia assigns to households both the physical properties of a housing unit, such as latitude and longitude, as well as properties about the household as a social unit, such as total household income, race and age of the head of the household, and household size. Our approach distinguishes household from housing unit and asserts that housing units are individuated by their residence functions and that a household realizes the housing unit's residence function by living there. In OMRSE, we define a household as *a human or collection of humans that occupies a housing unit by storing their possessions there and habitually sleeping there thereby participating in the realization of its residence*

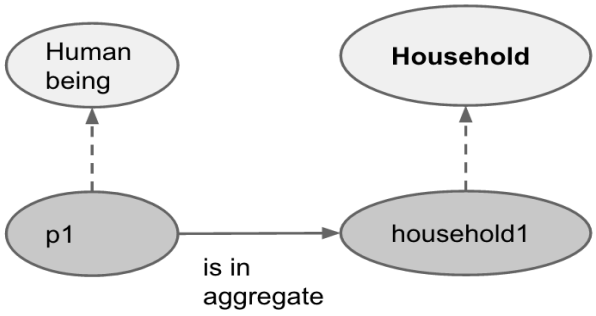

Fig 2. Graphical Model of RDF Specification of Household.

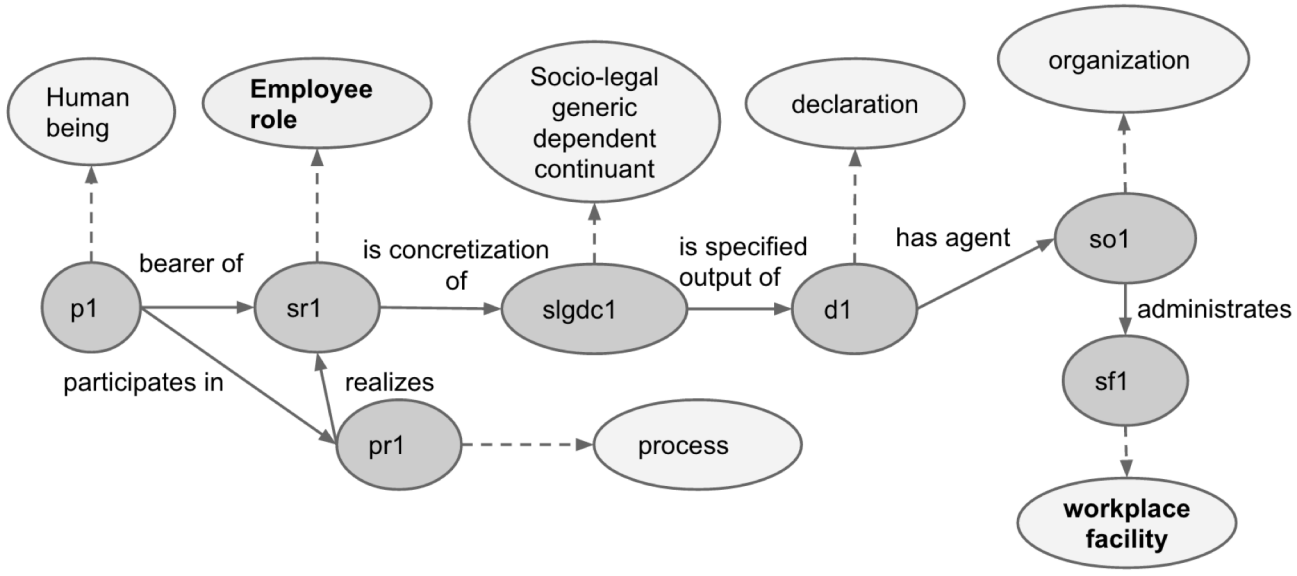

Fig 3. Graphical Model of RDF Specification of a Person's Relation to a Workplace.

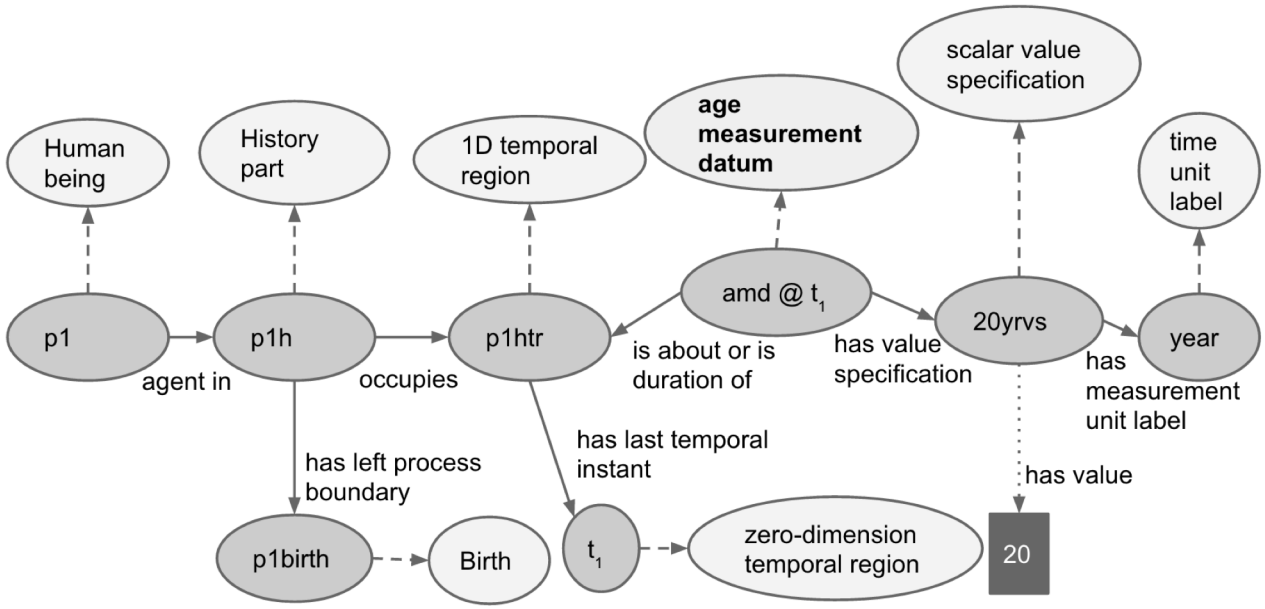

Fig 4. Graphical Model of RDF Specification of Age.

function, and add the following description logic equivalence statement:

household =def ('Homo sapiens' or 'collection of humans') and ('participates in' some (process and (realizes some 'residence function')))

where residence function is defined as *a function that inheres in a material entity and is realized by protecting persons and their possessions from weather and by some person or group of persons habitually sleeping in at least one site that is contained by that material entity*.

B. New OMRSE Classes

We created a total of 11 new classes in OMRSE to support the representation of synthetic ecosystems. Each class has a textual definition adapted from U.S. Census. One major adaptation of the definitions was to put them in Aristotelian form with the name of the direct superclass as part of the definition. Other adaptations were necessary to eliminate ambiguity and to reuse other defined ontology terms. OMRSE is a publicly-available resource at the following permanent URL: http://purl.obolibrary.org/obo/omrse.owl.

C. RDF Datasets and Queries

The Alachua county dataset comprised ~13M triples, and the Miami-Dade County dataset comprised 133M triples (Table 1). The population totals for both counties are slightly lower than the 2010 Census numbers on which the Synthia datasets were based. The reason is that we did not incorporate group quarters such as nursing homes and military barracks, which is future work.

The execution time for the SPARQL queries ranged from a few milliseconds to 41 seconds. The longest of these was the query that counted all workplaces with at least one employee who lives at home with at least one student.

TABLE I. SUMMARY STATISTICS FOR TWO COUNTY-BASED DATASETS

	Alachua	**Miami-Dade**
Triples	13,315,702	133,973,948
People	233,549	2,448,514
Schools	64	442
Workplaces	13,895	180,773
Housing Units	100,517	867,252
Average Household Size	2.32	2.82
Employees per workplace	8.05	6.13
Students per school	584	1070
Workplaces that overlap with a school	7895 (56.8%)	121,951 (67.5%)
Households with both an employee and a student	20,244 (20%)	255,614 (29.5%)

The housing unit totals for both counties match the 2010 Census numbers. The data show distinct differences, as expected, between Miami-Dade—a large urban county—and Alachua—a small county (in terms of population) where a large university is located. Miami-Dade has a larger household size

and school size, a greater percentage of workplaces with at least one employee that lives with at least one school student, and a greater percentage of households with at least one workplace employee and school student. By contrast, Alachua has a higher average workplace size, even when the University of Florida is excluded from consideration. These differences are likely to impact simulator results—Miami-Dade will often have a larger incidence and prevalence of infectious disease that is spread from person to person such as influenza in the absence of control measures. Infectious disease control measures designed to reduce school and workplace transmission—such as school closure, voluntary or imposed absenteeism from work, and vaccination of the school and / or workplace population—are likely to have a greater predicted effectiveness (and thus perhaps actual effectiveness) in Miami-Dade than Alachua.

D. *Availability of Materials*

All materials created for this paper—the graphical models (including additional ones not shown here), the SPARQL queries, and the OWL files with the entire datasets for Alachua and Miami-Dade counties—are freely available under a Creative Commons Attribution (CC BY 4.0) license at: http://tinyurl.com/syneco-queries.

IV. Discussion

We developed a Semantic Web and realism-based representation of the entire populations of two counties in Florida. We built SPARQL queries to assess differences between the two populations that are likely to influence disease transmission, as well as the results of experiments conducted using DTMs. The approach is generic and could be applied to any other synthetic ecosystem data, including for additional geographical regions. The queries are generic and could be applied to any additional county-based datasets (or datasets at other levels of geographical granularity such as Census tract) similarly transformed via our processes and representations.

We have demonstrated the feasibility of using Semantic Web technologies for representing entire populations, and in particular for representing synthetic ecosystems for use in AB-DTMs. Additionally, through additions to OMRSE and the creation of RDF synthetic datasets, we have developed some of the resources necessary to transform other U.S. Census data into Semantic Web representations. In so doing, we have made explicit much of the semantics that are implicit in those data and the synthetic ecosystems that are based on them. It is our conjecture for future work that the explicit semantics improve the ease with which synthetic ecosystems can be expanded to incorporate additional biological, social, and abiotic ecosystem elements.

Although we developed this work in the context of agent-based DTMs, this resource and approach could also be leveraged for social network analysis due to the graph-based nature of RDF. For example, one could construct queries for finding hubs in the network and people or places that a set of people have in common. Furthermore, DTMs are increasingly taking into account social networks as part of the synthetic ecosystem itself (for example, see Frias-Martinez et al. [20]). Network-based approaches and graph representations such as our RDF-based one here are more extensible and suitable for representing these networks.

Future work includes expanding the specification to include data related to group quarters, which will require additional ontological analysis and ontology development.

Acknowledgments

This work was supported by award UL1TR001427 from the National Center for Advancing Translational Sciences (NCATS) and award U24GM110707 from the National Institute for General Medical Science (NIGMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of NCATS, NIGMS, or the NIH.

References

[1] S. T. Brown, J. H. Tai, R. R. Bailey, P. C. Cooley, W. D. Wheaton, M. A. Potter, et al., "Would school closure for the 2009 H1N1 influenza epidemic have been worth the cost?: a computational simulation of Pennsylvania," BMC Pub. Health, vol. 11, p. 353, 2011.

[2] M. E. Halloran, N. M. Ferguson, S. Eubank, I. M. Longini, D. A. Cummings, B. Lewis, et al., "Modeling targeted layered containment of an influenza pandemic in the United States," Proceedings of the National Academy of Sciences of the United States of America, vol. 105(12), pp. 4639-4644, 2008.

[3] I. M. Longini, A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. Cummings, and M. E. Halloran, "Containing pandemic influenza at the source," Science, vol. 309(5737), pp. 1083-1087, 2005.

[4] C. Siettos, C. Anastassopoulou, L. Russo, C. Grigoras, and E. Mylonakis, "Modeling the 2014 ebola virus epidemic – agent-based simulations, temporal analysis and future predictions for Liberia and Sierra Leone," PLOS Currents Outbreaks, Edition 1, 2015.

[5] J. J. Grefenstette, S. T. Brown, R. Rosenfeld, J. DePasse, N. Stone, P. C. Cooley, et al., "FRED (a Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations," BMC Pub. Health, vol. 13, p. 940. 2013.

[6] W. D. Wheaton, J. C. Cajka, B. M. Chasteen, D. K. Wagener, P. C. Cooley, L. Ganapathi, et al., 2009. "Synthesized population databases: a US geospatial database for agent-based models," Methods Report, RTI Press, 2009(10), p. 905.

[7] MIDAS Informatics Services Group, "Synthetic Populations and Ecosystems of the World," 2016. http://data.olympus.psc.edu/syneco/spew_documentation.pdf

[8] MIDAS, "Synthetic Populations and Ecosystems," 2014. http://www.epimodels.org/drupal-new/?q=node/112

[9] W3C, "RDF Current Status," https://www.w3.org/standards/techs/rdf#w3c_all. Last accessed 06/20/2016.

[10] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, et al., "TWC LOGD: a portal for linked open government data ecosystems," Journal of Web Semantics, vol. 9(3), pp. 1–11, 2011.

[11] W. R. Hogan, S. Garimalla, and S. A. Tariq, "Representing the reality underlying demographic data," In Proceedings of the International Conference on Biomedical Ontology, pp. 147-152, Buffalo, NY: International Conference on Biomedical Ontology, 2011

[12] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, et al., "The Obo Foundry: coordinated evolution of ontologies to support biomedical data integration," Nature Biotechnology, vol. 25(11), p. 1251, 2007.

[13] P. Grenon and B. Smith, "Snap and span: towards dynamic spatial ontology," Spatial Cognition and Computation, vol. 4(1), pp. 69-104, 2004

[14] S. Federhen, "The NCBI taxonomy database," Nucleic Acids Research, vol. 40(D1), pp. D136-D43, 2012.

[15] W. Ceusters, Ed. “An information artifact ontology perspective on data collections and associated representational artifacts,” MIE, 2012.

[16] M. B. Almeida, L. Slaughter, and M. Brochhausen, Eds. “Towards an ontology of document acts: introducing a document act template for healthcare,” In On the Move to Meaningful Internet Systems: OTM 2012 Workshops, Rome, Italy: Springer, 2012, pp. 420-425.

[17] W3C, “OWL 2 Web Ontology Language Document Overview (Second Edition),” 2012. https://www.w3.org/TR/owl2-overview/. Last accessed 06/20/2016.

[18] M. Brochhausen, W. R. Hogan, J. Levander, S. T. Brown, N. Millet, J. Hanna, et al., 2014. “A novel representation of terms related to infectious disease epidemiology for epidemic modeling: the Apollo Structured Vocabulary and pre-existing representations,” In Proceedings of the International Conference on Biomedical Ontology, Houston, Texas: CEUR Workshop, W.R. Hogan, S. Arabandi, and M. Brochausen, Eds. 2014, pp. 21-26.

[19] R. R. Brinkman, M. Courtot, D. Derom, J. M. Fostel, Y. He, P. Lord, et al., “Modeling biomedical experimental processes with OBI,” Journal of Biomed Semantics, vol. 1 (Suppl 1), p. S7, 2010.

[20] E. Frias-Martinez, et al., “An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information,” Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011.