The Genomic Epidemiology Ontology and GEEM Ontology Reusability Platform

Damion DOOLEY[a,1], Emma GRIFFITHS[b], Gurinder GOSAL[a], Fiona BRINKMAN[b], and William HSIAO[a,b,c]
[a] *Department of Pathology and Laboratory Medicine, UBC, Vancouver, Canada*
[b] *Department of Molecular Biology and Biochemistry, SFU, Burnaby, Canada*
[c] *BCCDC Public Health Laboratory, Vancouver, Canada*

Abstract. There is an increasing awareness within private and public organizations that ontologies (globally accessible and uniquely identified terms that have both natural language definitions and logic relations which can be queried and reasoned over by computers) are useful in solving interoperability quagmires between data silos and the add-hoc data dictionaries that describe them. However, the complexity of implementing evolving ontologies in content management and federated data querying applications is formidable. The Genomic Epidemiology Entity Mart (GEEM) web platform is a proof-of-concept web portal designed to provide non-ontologist users with an ontology-driven interface for examining data standards related to genomic sequence repository records. GEEM provides web forms that show labels and allowed-values for easy review. It also provides software developers with downloadable specifications in JSON and other data formats that can be used without the need for ontology expertise. New systems can adopt ontology-driven standards specifications from the start, and the same specifications can be used to facilitate and validate the conversion of legacy data.

Keywords. Ontology, controlled vocabulary, integration, portal, biomedical metadata, biomedical research, epidemiology

1. Introduction

A survey of data sharing challenges in the public health domain[1] acknowledges a spectrum of technical to social barriers that need to be resolved for data integration projects to succeed. The GEEM project explores the potential for ontology technology to resolve data integration issues caused by incompatibilities in the use of numeric data types[2], data dictionaries, and categorical terms that do not automatically map to each other across systems. For example, concepts underlying terms such as “date of birth”, “age” or “age at time of specimen extraction” are shared by a number of databases and data standards. Providing a system for detailing the particular field labels, numeric and categorical choices and data type variations that different 3rd party standards or custom databases have would lay the necessary foundation for developing semi-automated data conversion tools. Accomplishing this using ontology technology will require standards for representing basic variables, as well as an emerging library of ontology-driven composite entities such as a patient’s symptoms on some date. There are a number of

[1] Damion Dooley, Department of Pathology and Laboratory Medicine, University of British Columbia, BCCDC Site, 655 West 12th Avenue, Vancouver, BC, Canada; E-mail: damion.dooley@bccdc.ca.

projects that provide ontology-driven user interfaces for managing or searching their platform[3,4,5], however these systems involve more complex data models which do not provide easy ways to interact with 3rd party standards.

The GEEM project is testing an ontology approach to describe standards by representing a number of test case standards such as the United States National Center for Biotechnology Information (NCBI) BioProject, BioSample and Antibiogram submission specifications[6], the related National Institute of Allergy and Infectious Disease GSCID-BRC core BioProject and core BioSample specifications, the Genomic Standards Consortium Minimum Information about any (x) Sequence (MIxS) standard's clinical specimens field package[7], and a sequence repository accreditation specification in development for recording food borne pathogen specimen contextual data. A short use-case example of existing data conversion problems is the variation by parties like NCBI and industry in use of units like µg/ml and mg/L to record minimum inhibitory concentration (MIC) drug dosages. NCBI requests mg/L to avoid an occasional problem of submitters erroneously having the "µ" micro character transformed into "m" during data export preparation, yielding mg/ml. Variations on unit spelling add complexity too, like "litre" and "liter". Ontology-marked-up data prevents these conversion issues by referencing a global unique identifier for each unit, e.g. http://purl.obolibrary.org/obo/UO_0000273 = "mg/L".

The first phase of the GEEM project, nearing completion, provides an interface for users to browse available specifications and their parts, see them rendered in example web forms which are nevertheless fully functional, and to download the specifications in a variety of formats – in full JSON or YAML format, or a further simplified version thereof. Figure 1 shows the GEEM experimental prototype, currently under active development at http://genepio.org/geem/.

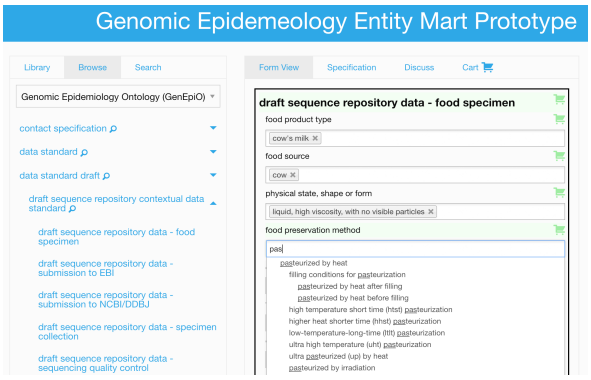

Figure 1. The GEEM website allows data standards to be listed and reviewed in interactive forms.

Browse and Search tabs explore the contents of a selected ontology, currently limited to those described in the Genomic Epidemiology Ontology (GenEpiO)[8], an application ontology for describing the use of pathogen genomes and their clinical, environmental or foodborne context in epidemiological surveillance and investigation. Form View and Specification tabs enable interactive review and downloading of selected items. The upcoming Discuss and Cart tabs will provide community discussion about a particular term, and will allow different specifications to be brought together into one downloadable and refreshable package. The GEEM system is designed to insulate standards reviewers and implementers from the current complexity

involved in developing and maintaining ontologies. They do not need to know about ontology technology other than the attractiveness of global identifiers, hierarchic term facets, and multilingual capability.

2. Approach

GEEM takes as input an ontology that combines terms and relationships from other ontologies to describe the interdisciplinary domains of knowledge that a standard involves, as well as a specification for what precisely a standard requires for its fields from the ontology. As a test case we have used GenEpiO, which the Hsiao Lab initiated in 2015 as part of the Integrated Rapid Infectious Disease Analysis project (www.irida.ca). GenEpiO draws upon the OBOFoundry family of ontologies[9], collectively covering the domains of symptoms, diseases, anatomy, food, geography, environment, and taxonomy, among others, to describe the standards it seeks to fulfill. Figure 2 shows a portion of GenEpiO standards entities that GEEM renders as forms and specifications. Note that GEEM does not commit to how standardized data is ultimately stored; this is an implementation detail that will vary between institutions.

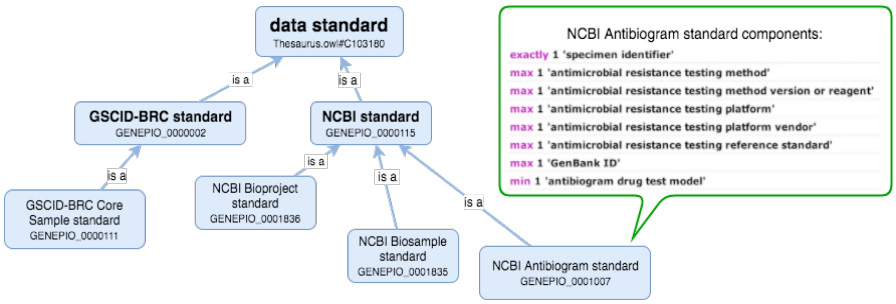

Figure 2: Example representations of standards within the GenEpiO application ontology. Each standard is a collection of either individual field or composite field entities.

The W3C OWL specification inherited XML data types, allowing entities to have URI, integer, decimal, integer, string, and date-time values as exemplified in the purple boxes in Figure 3. GEEM needs to describe the expected data type(s) that a standardized measurable or variable has, but Ontology for Biomedical Investigations (OBI) lacks such a relationship - claims can't be made about an entity's data type independently of an instance of a stored value. Consequently we added a "has primitive data type" data property relation, as shown in Figure 3, that points directly to an OWL data type, and which allows a subclass of similar variables/inputs/outputs to inherit the same. This enables constraints on an associated value, if numeric, or its length, or character pattern, if a string. GEEM also has a "categorical tree specification" class to hold categorical variables that provide hierarchies of choices. A unit (meters, seconds, etc.) can be tacked onto a scalar datum by way of the "has measurement unit label" relation.

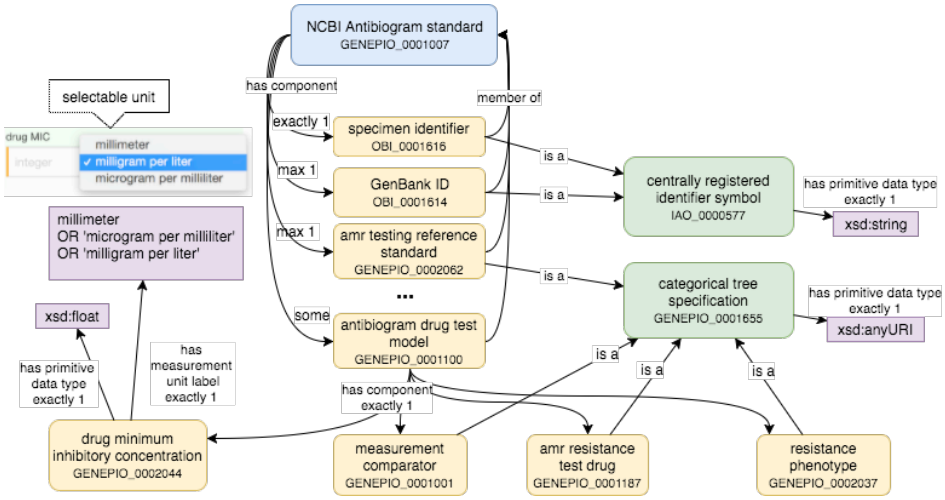

Figure 3. The GenEpiO schema for NCBI's Antibiogram standard draws upon several OBOFoundry ontologies including the Unit Ontology (UO) and Ontology for Biomedical Investigations (OBI).

More work needs to be done to harmonize approaches to units, as there are at least three unit ontologies in use: Quantities, Units, Dimensions and Data Types Ontologies[10] (QUDT), Ontology of units of Measure and related concepts[2] (OM), and the Units Ontology[11] (UO). QUDT and OM illustrate how units are a microcosm of ontological complexity that are under-utilized if left as atomic terms. They enable a compound unit to be decomposed into specifications for its numerator and divisor, which can then enable unit analysis or unit conversion, e.g. OM's "Compound units", or QUDT's "Quantity Dimensions". Ideally a datum's preferred unit and scale are stated (e.g. Celsius vs. Kelvin) to encourage migration of data to the preferred format.

Currently GEEM relies entirely on the pre-processing of an ontology file and its imports into a single JSON data structure that can then be loaded via AJAX and interpreted by browser-based javascript menu and form rendering scripts. The python "jsonimo.py" script applies SPARQL queries to a memory resident GenEpiO data structure created by python's rdflib package, and processes the results into a fairly flat file of entity representations as shown in step 2 of Figure 4.

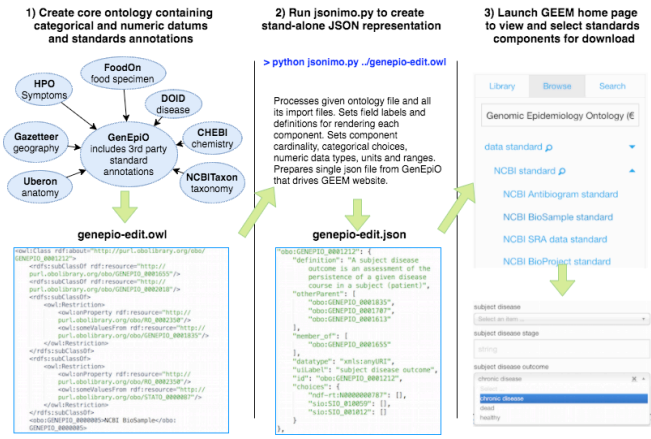

Figure 4. The GEEM website content is generated entirely from a single JSON file.

Other core ontologies can similarly have their JSON file representations generated for GEEM to use. The GEEM website is driven by zurb Foundation 6.0 and jQuery but it would be attractive to have other form rendering platforms active to demonstrate a non-competitive open-source potential. The upcoming introduction of user accounts will require server-based services to persist user-defined packages of specifications, but a stand-alone form rendering script (http://genepio.org/geem/form.html) demonstrates how lightweight this technology can be. In the near future, GEEM-enabled Microsoft Excel spreadsheet templates will be provided that have fully functional data entry.

The GEEM project's initial aim is to provide adequately detailed and standards-compliant specifications that can fully describe specimen metadata for selected food borne (*Salmonella*, *E. coli*, *Listeria*) and other (*M. tuberculosis*) pathogens to support infectious disease analysis without any further vocabulary needed from other sources. This does depend on quality curating of the underlying ontologies so that missing or ill-defined terms are quickly remedied. After gaining consensus with the OBI and other ontology communities about how 3rd party data standards should be represented, we expect a strong foundation on which to study the remaining data conversion challenge.

Acknowledgements

The authors would like to express their appreciation for feedback from JOWO 2017 ODLS workshop participants. This study is supported by Genome Canada / CIHR Bioinformatics and Computational Biology (BCB) grant (254EPI/ BOP-149425) and Genome BC grant (B15DSI).

References

[1] van Panhuis, W. G., et al.. A systematic review of barriers to data sharing in public health. BMC Public Health **14** (2014), 1144.

[2] M. van Assem, H. Rijgersberg, and J. Top, Ontology of units of measure and related concepts. Semantic Web **4 no. 1** (2013), 3-13.

[3] Lozano-Rubí R et. al. OWLing Clinical Data Repositories With the Ontology Web Language. Eysenbach G, ed. JMIR Medical Informatics. **2.2** (2014) e14.

[4] Haendel M et. al. eagle-i: ontology-driven federated search and data entry tools for discovering biomedical research resources. Proceedings of the 4th International Conference on Biomedical Ontology; 4th International Conference on Biomedical Ontology (2013)

[5] González-Beltrán, Alejandra, et al. linkedISA: semantic representation of ISA-Tab experimental metadata. BMC bioinformatics **15.14** (2014) S4.

[6] Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. **40** (2012) D57–63.

[7] Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol **29** (2011) 415–420.

[8] Griffiths Emma et al. Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance, Frontiers in Microbiology **8** (2017), 1068.

[9] Smith, Barry et al. The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nature biotechnology **25.11** (2007) 1251.

[10] R. Hodgson and P. J. Keller. QUDT-quantities, units, dimensions and data types in OWL and XML. Online (2011), http://www.qudt.org

[11] Gkoutos, Georgios V., Paul N. Schofield, and Robert Hoehndorf. The Units Ontology: A Tool for Integrating Units of Measurement in Science. Database: The Journal of Biological Databases and Curation 2012 (2012), bas033.