Malaria study data integration and information retrieval based on OBO Foundry ontologies

Jie Zheng, JaShon Cade, Brian Brunk, David S. Roos, Christian J. Stoeckert Jr.
EuPath Bioinformatics Resource Center
University of Pennsylvania
Philadelphia, PA, USA

San Emmanuel James, Emmanuel Arinaitwe
Infectious Diseases Research Collaboration
Kampala, Uganda

Bryan Greenhouse, Grant Dorsey
Department of Medicine
University of California San Francisco
San Francisco, CA, USA

Steven A. Sullivan, Jane M. Carlton
Center for Genomics & Systems Biology
Department of Biology
New York University
New York, NY, USA

Gabriel Carrasco-Escobar, Dionicia Gamboa
Universidad Peruana Cayetano Heredia
Lima, Peru

Paula Maguina-Mercedes, Joseph M. Vinetz
Division of Infectious Diseases
University of California San Diego
La Jolla, CA, USA

Abstract— **The International Centers of Excellence in Malaria Research (ICEMR) projects involve studies to understand the epidemiology and transmission patterns of malaria in different geographic regions. Two major challenges of integrating data across these projects are: (1) standardization of highly heterogeneous epidemiologic data collected by various ICEMR projects; (2) provision of user-friendly search strategies to identify and retrieve information of interest from the very complex ICEMR data. We pursued an ontology-based strategy to address these challenges. We utilized and contributed to the Open Biological and Biomedical Ontologies to generate a consistent semantic representation of three different ICEMR data dictionaries that included ontology term mappings to data fields and allowed values. This semantic representation of ICEMR data served to guide data loading into a relational database and presentation of the data on web pages in the form of search filters that reveal relationships specified in the ontology and the structure of the underlying data. This effort resulted in the ability to use a common logic for storing and display of data on study participants, their clinical visits, and epidemiological information on their living conditions (dwelling) and geographic location. Users of the Plasmodium Genomics Resource, PlasmoDB, accessing the ICEMR data will be able to search for participants based on environmental factors such as type of dwelling, location or mosquito biting rate, characteristics such as age at enrollment, relevant genotypes or gender and visit data such as laboratory findings, diagnoses, malaria medications, symptoms, and other factors.**

Keywords—standardizing data dictionaries, OBO Foundry, PlasmoDB, ICEMR

I. INTRODUCTION

The ICEMR program is a global network of 10 independent research centers created to improve understanding of the epidemiology and transmission patterns of malaria in different geographic regions [1]. Integrating data generated by these Centers into the Plasmodium Genomics Resource (PlasmoDB) [2], a component of the Eukaryotic Pathogen Bioinformatics Resource Center (EuPath BRC), provides web-enabled access to ICEMR project members, and ultimately the broader international research community. Common data collected across all ICEMR projects are represented in Figure 1. However, data produced by the various ICEMR projects is heterogeneous with respect to origin, type of data, format, and spatio-temporal scale. The main challenges of sharing and
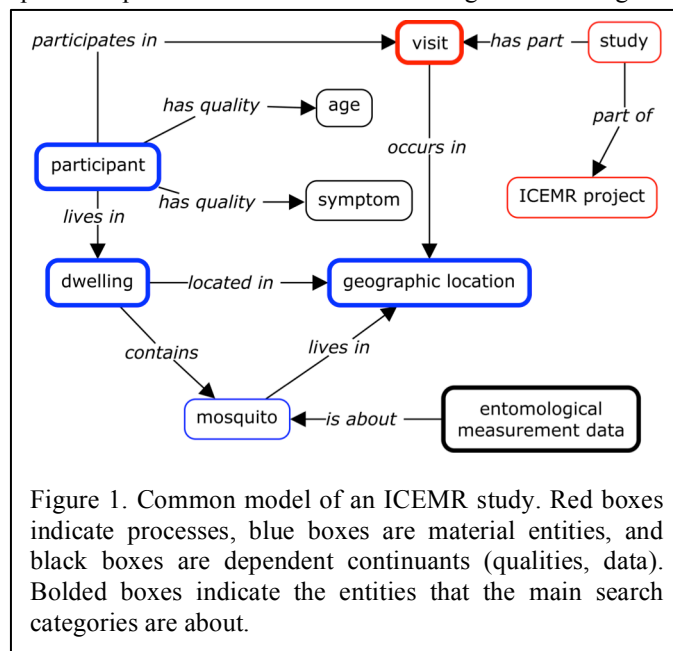

Figure 1. Common model of an ICEMR study. Red boxes indicate processes, blue boxes are material entities, and black boxes are dependent continuants (qualities, data). Bolded boxes indicate the entities that the main search categories are about.

integration of ICEMR data include standardizing the complex and heterogeneous data for consistent representation and providing a user-friendly interface for easy exploration of the data for constructing searches..

Ontologies play a crucial role in heterogeneous data integration by supporting consistent data representation and providing a semantic framework to reveal the relationships between data thereby facilitating information retrieval and new knowledge discovery [3]. We made use of the Open Biological and Biomedical Ontologies (OBO) Foundry [4] which promotes interoperable ontologies and provides a listing of ontologies seeking to follow Foundry principles. These ontologies were used to provide a common understanding of what the information collected according to different ICEMR data dictionaries and case record forms was about. The OBO-based mappings were useful for guiding data loading and queries but were not directly usable for providing intuitive display of the available data on search forms. These were combined in a EuPath application ontology. Using WebProtege [5], we created an ICEMR terminology to organize the classes of data, create top-level categories, and re-label terms according to user preference while still maintaining the OBO IRIs where applicable to preserve the semantic underpinnings. The result was a linked OBO-based application ontology and web display terminology to provide interoperability and intuitive access to the datasets based on different data dictionaries.

II. METHODS

A. ICEMR data and data dictionaries

Multiple ICEMR projects have provided data for inclusion in PlasmoDB. Each ICEMR project has provided data dictionaries covering all data variables and values required for interpreting the associated data. By data dictionary, we mean a list of terms with definitions and specification of data variables, data types, format of data, and allowed values (including controlled vocabulary values). Data dictionaries are used in data exchanges among ICEMR projects and sharing with different repositories. However, data dictionaries from the different ICEMR projects generally look very different from each other in terms of type and quantity of content.

B. Consistent representation of ICEMR data

To standardize the data dictionaries from different ICEMR projects, the variables and controlled vocabulary values were mapped to OBO ontologies. These included the Ontology for Biomedical Investigations (OBI) [6], Phenotype qualities (PATO) [7], Ontology for General Medical Science (OGMS) [8], Environmental Ontology (EnVO) [9], Disease Ontology (DO) [10], Drug Ontology (DRON) [11], Infectious Disease Ontology (IDO) [12], Human Phenotype Ontology (HP) [13], Information Artifact Ontology (IAO) [14], Ontology for Biobanking (OBIB) [15], and Symptom Ontology (SYMP) [16]. The mapping of terms specified in the data dictionaries to OBO ontologies was performed using the BioPortal annotator web services [17]. The annotator service can accurately (>95%) tag text with ontology terms. However, ontologies in the annotator might not be the latest version since these need to go through an indexing process before being added to the annotator. For terms where mappings were not found using the

annotator, the BioPortal search web services [18] were used. Both annotator and search results were reviewed manually.

Consistent representation of ICEMR data was achieved once the variables and values in the different ICEMR data dictionaries were either mapped to existing ontology terms or new ontology terms were created for that purpose. New ontology terms were created using two approaches.

a) If the terms were general and in a domain which have been covered by an OBO ontology, they were submitted to the relevant ontology via its issue tracker to be added in by the ontology developers. For example, disease terms were submitted to the DO tracker and terms related to the environment were submitted to the EnVO tracker.

b) If the terms were specific to the ICEMR projects, they were added in the Eupath ontology. The Eupath ontology is an application ontology developed for providing terms to annotate data in the EuPath BRC. The EuPath ontology was built based on OBI with integration of other OBO ontologies such as PATO, OGMS, DO, etc. when needed.

C. Organization of ICEMR data dictionary variables for guiding searches of ICEMR data

The ontological mapping of data dictionary variables provides semantic clarity of types. However, organization according to term types (e.g., processes, material entities, qualities, etc.) does not necessarily provide intuitive listing on web sites for mining the data. As illustrated in Figure 1, the five main types of interest are 'participants', 'dwellings', (clinical) 'visits', 'entomological measurements' and 'geographic location'. Therefore, we organized the data dictionary variables into categories based on their relation to these types. Within each category, the data dictionary variables are grouped based on the mapped OBO ontology terms. For example, 'height', 'weight', and 'temperature' (measurement data) are grouped together in the 'physical examination' category (which in turn is placed in the 'visit' category). The outcome of categorization of the variables from the multiple ICEMR data dictionaries is the ICEMR terminology and is the basis for displaying search parameters of this data on the PlasmoDB website. The ICEMR terminology is represented in the OWL format containing only 'is a' relations enabling visualization of the ICEMR data dictionary hierarchy organization using ontology editors. WebProtege is a web-based collaborative ontology development platform and provides a means for domain experts to review and post comments on terms [5]. We uploaded the ICEMR terminology to WebProtege and used it for collaboratively reviewing both the organization of the ICEMR terminology and the labels of terms to be displayed on the PlasmoDB web site before loading the ICEMR data into the database. This approach ensured that the data was correctly displayed on PlasmoDB for each ICEMR project. For the ICEMR terminology, we specified display labels using the rdfs:label annotation property as they are the default term labels rendered on WebProtege. In addition, we used annotation properties to specify ontological names, definitions, whether the term was an organizing category or a variable. If the term corresponded to a data dictionary variable, then annotation properties were also used for the original variable name in the data dictionary and source,

the mapped ontology term, and the ontological definition. The common display labels in the ICEMR terminology were agreed upon by the contributing ICEMR projects. Each contributing ICEMR project had variables unique to that project. Therefore, the application of the ICEMR terminology for organization of each ICEMR data dictionary resulted in different but still consistent outputs. The application of the ICEMR terminologies to the different projects can be viewed at the WebProtege site (http://webprotege.stanford.edu/) as "ICEMR Amazonia", "ICEMR Indian", and "ICEMR PRISM" (Uganda ICEMR project).

III. RESULTS

A. ICEMR data and data dictionaries

Longitudinal data from three ICEMR projects with studies in Uganda, India, and Amazonia were submitted for inclusion in PlasmoDB. Data and data dictionaries from the Uganda and Indian ICEMR projects were provided in English whereas data and the data dictionary from the Amazonia ICEMR project were in Spanish. The Amazonia ICEMR project also provided a translated data dictionary in English. All three ICEMR projects provided participant data, dwelling data on participants, and participant-associated clinical visit data. The Uganda ICEMR project also submitted entomological measurement data.

The Amazonia ICEMR data dictionary included 84 variables and 179 controlled values for 26 variables. The Indian ICEMR data dictionary contained 118 variables with 149 controlled values for 32 variables. The Uganda ICEMR data dictionary contained 121 different kinds of variables and 481 controlled values for 21 variables.

B. Ontology term mapping

Variables and values specified in the ICEMR data dictionaries were mapped to 10 different OBO Foundry ontologies (listed in the Methods). Table 1 lists the mapping results for each ICEMR project. A total of 209 new terms were added to the EuPath ontology for unmapped ICEMR variables. The EuPath ontology can be viewed on the WebProtege site (http://webprotege.stanford.edu/) as the "EuPath ontology" project.

Table 1. Summary of mapped ontology terms

ICEMR Project	Variables	OBO Ontologies	EuPath Ontology
Amazonia	84	15	69
India	118	31	87
Uganda	121	17	104

Data dictionary variables from the different ICEMR projects referring to the same thing were often different. For example, "edad" in the Amazonia ICEMR data dictionary, "age_en" in the Indian ICEMR data dictionary, and "age" in the Uganda ICEMR data dictionary all refer to participant age at the time of enrollment and mapped to the ontology term EUPATH_0000120: 'age since birth at time of enrollment'. As another example of the encountered heterogeneity, Table 2

shows a sampling of mapping between symptom related variables to ontology terms.

Ontology term mapping was also performed on the controlled values of variables. 413 controlled values used in the Uganda ICEMR data were mapped to OBO ontology terms. The remaining 68 unmapped terms were added into the EuPath ontology. Few corresponding ontology terms were found for the controlled values in the Amazonia and Indian ICEMR data (14 for Amazonia and 5 for Indian, respectively). For those values without mapped ontology terms, we have created standardized labels and will add the terms to either OBO ontologies or EuPath ontology as described in the Methods.

After ontology term mapping and standardization of value labels across data from multiple ICEMR projects, we generated (data dictionary to standardized) term mapping files for each ICEMR. These mapping files were used in the ICEMR project data loading process and enabled consistent data representation in the PlasmoDB database.

Table 2. Ontology mapping of symptom related variables

Data dictionary	Ontology term ID	Ontology term label	ICEMR display name
abdominalpain	HP_0002027	Abdominal pain	Abdominal pain
apainduration	EUPATH_0000154	duration of abdominal pain	Abdominal pain duration
Anorexia	SYMP_0000523	anorexia	Anorexia
aduration	EUPATH_0000155	duration of anorexia	Anorexia duration
Cough	SYMP_0000614	cough	Cough
cduration	EUPATH_0000156	duration of cough	Cough duration
Diarrhea	DOID_13250	diarrhea	Diarrhea
dduration	EUPATH_0000157	duration of diarrhea	Diarrhea duration
Fatigue	SYMP_0019177	fatigue	Fatigue
fmduration	EUPATH_0000158	duration of fatigue	Fatigue duration
febrile	EUPATH_0000097	febrile	Febrile
fever	EUPATH_0000100	subjective fever	Fever (subjective)
Headache	HP_0002315	Headache	Headache
hduration	EUPATH_0000159	duration of headache	Headache duration
Jaundice	HP_0000952	Jaundice	Jaundice
jduration	EUPATH_0000160	duration of jaundice	Jaundice duration
jointpains	SYMP_0000064	joint pain	Joint pains
djointpains	EUPATH_0000161	duration of joint pains	Joint pains duration
muscleaches	EUPATH_0000252	Muscle aches	Muscle aches
mduration	EUPATH_0000162	duration of muscle aches	Muscle aches duration
rfa	OGMS_0000015	clinical history	Other medical complaint
seizure	SYMP_0000124	seizure	Seizures
sduration	EUPATH_0000163	duration of seizures	Seizures duration
fduration	EUPATH_0000164	duration of subjective fever	Subjective fever duration
Vomiting	HP_0002013	Vomiting	Vomiting
vduration	EUPATH_0000165	duration of vomiting	Vomiting duration

C. Organization of terms for search filters and exploration of data

For each ICEMR project, around 100 different variables can be used to search and retrieve the data. As indicated in the Introduction, malaria researchers are interested in mining the data for insights about the connections between study participants, their living conditions (dwelling), their health status (clinical visit), their geographic location and exposure to mosquitos (entomological measurement data). We assigned the variables to these five categories based on their mapped

ontology terms taking into account whether they were a subclass of or having a logical connection to the categories. With the exception of geographic location, each category had around 20 different variables that required further grouping to provide intuitive access to the data for end users. Further grouping was made based on the ontological understanding of data. For example, height, weight, and temperature data are all generated by physical examination. Thus, a new class of data OGMS_0000083: 'physical examination' was added under category 'visit'. Using this approach, around 5 different subtypes were created under each category (except 'geographic location'). For example, in addition to 'physical examination', 'medication', 'diagnosis', 'symptoms', 'laboratory findings', 'visit type' and 'visit details' were added as subtypes of the category 'visit'.

Term labels used in an ontology are typically chosen for ontological clarity and can be quite long. As a result, such labels are often not user-friendly or practical for providing searches on web sites like PlasmoDB. Alternative display names were therefore generated for ontology terms. For example, the display name 'Age at time of enrollment' is used for ontology term EUPATH_0000120: 'age since birth at time of enrollment'.

Figure 2 shows the organization of variables that will be displayed on the website in the three ICEMR projects discussed here using Protégé, an OWL editor [19]. Among the

different ICEMR data are found common categories but also some categories specific to individual projects. Therefore, each ICEMR project has its own representation of the ICEMR terminology used as web site search filters to explore its data. The application of this approach for the Uganda ICEMR project is shown in Figure 3. The applications for the other ICEMRs will be very similar and therefore users familiar with one ICEMR search will also find the other ICEMR searches to be familiar. Furthermore, the common display and underlying ontology mappings provide the opportunity for future cross ICEMR searches.

IV. DISCUSSION/ CONCLUSIONS

Related but different semantic approaches were used to address the dual challenges of standardizing data dictionaries across projects and generating user-friendly displays to search and explore the associated data.

Our approach for standardization is to relate all variables and associated values to terms from interoperable ontologies listed at the OBO Foundry. OBO Foundry ontologies provide the benefit of wide coverage but can also be selectively imported to create an application ontology such as the EuPath ontology. When existing terms were not available for mapping, new ones were created for introduction into the source ontologies or just placed in the application ontology. The use of the Basic Formal Ontology (BFO) [20] by the EuPath ontology as its upper level greatly facilitated the task of
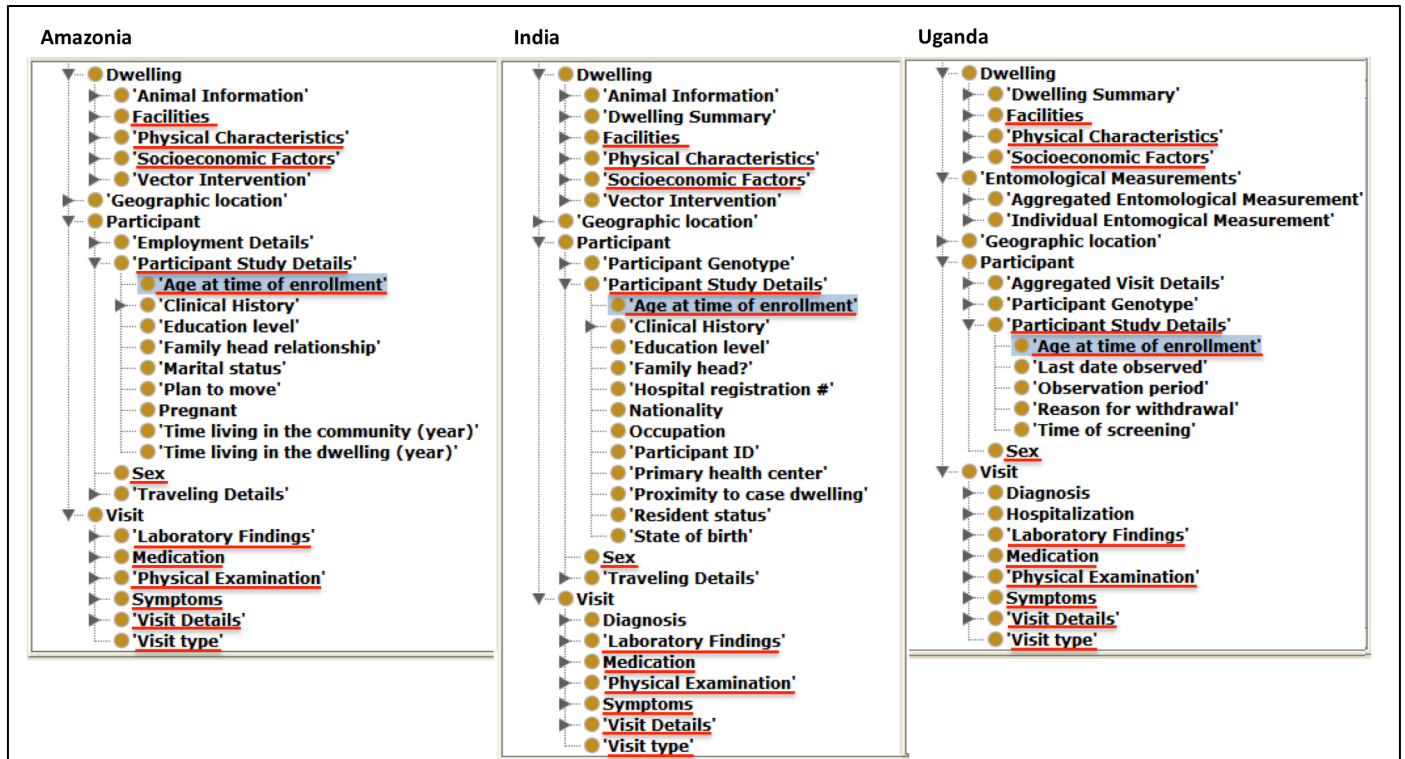

Figure 2. Standardized representation of variables from the Amazonia (left), India (middle), and Uganda (right) ICEMR data dictionaries for web display. Highlighted is an example of variable common to all three, 'Age at time of enrollment', which is placed under 'Participant Study Details' along with variables that are common to only two (e.g., 'Clinical History') or unique (e.g., 'Reason for withdrawal'). Other categories and variables common to all three ICEMRs are underlined in red.

standardization across projects. BFO models reality rather than data models and helps interpret when variables and values are about the same processes, material entities, and measurements. However, the ontologic semantic organization did not directly translate well to web site displays for exploring relationships between study participants, their living conditions, and data gathered at clinical visits to understand malaria epidemiology. Instead, categorical organization was better suited for web display.
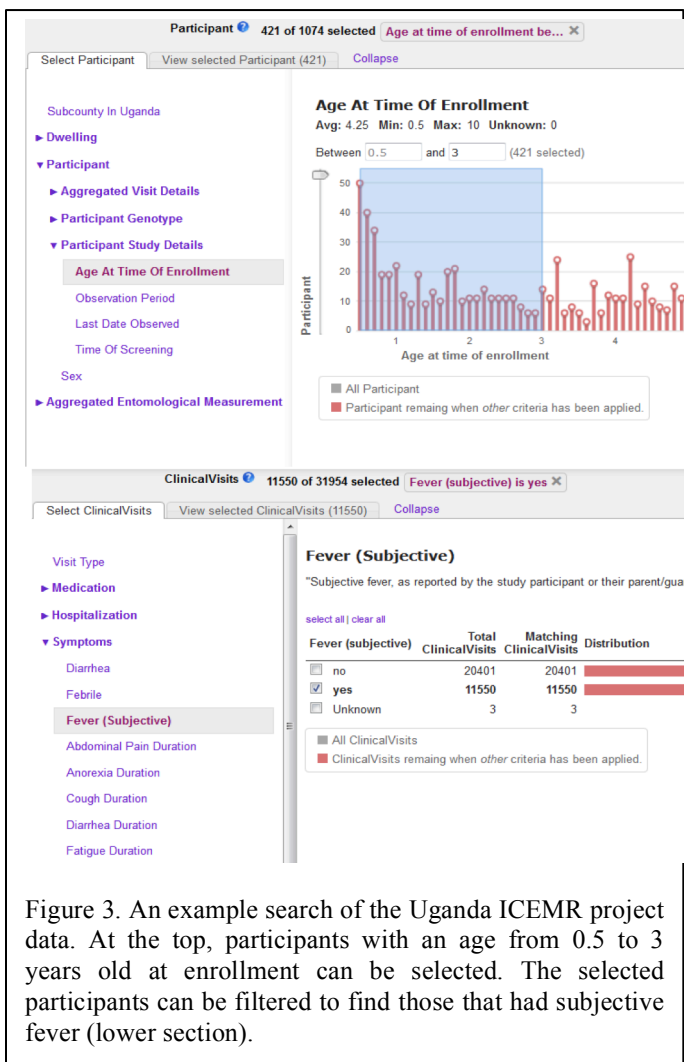

Figure 3. An example search of the Uganda ICEMR project data. At the top, participants with an age from 0.5 to 3 years old at enrollment can be selected. The selected participants can be filtered to find those that had subjective fever (lower section).

An ICEMR terminology was created for the purpose of web display to organize the standardized variables according to ways that users are expected to browse them. The ICEMR terminology also takes into account the need for shortened names on a web form. Underlying all the terms however is their basis for understanding through mapping to OBO / EuPath ontology terms.

The separation of web display and variable standardization provides for flexibility in providing different emphases in data browsing while maintaining the same underlying semantics. The overall approach has allowed us to achieve the goal of providing a common system with consistent representation for the three currently participating ICEMR projects. It also provides a flexible existing system for introducing data from other ICEMR projects or other studies of the same type.

REFERENCES

[1] J. B. Gutierrez, O. S. Harb, J. Zheng, D. J. Tisch, E. D. Charlebois, C. J. Stoeckert, et al., "A framework for global collaborative data management for malaria research," Am. J. Trop. Med. Hyg. vol. 93 no. 3 Suppl., pp. 124-32, September 2015.

[2] C. Aurrecoechea, J. Brestelli, B. P. Brunk, J. Dommer, S. Fischer, B. Gajria, et al., "PlasmoDB: a functional genomic database for malaria parasites," Nucleic Acids Res. vol. 37, pp. D539-43, January 2009.

[3] V. G. Dugan, S. J. Emrich, G. I. Giraldo-Calderón, O. S. Harb, R. M. Newman, B. E. Pickett, et al, "Standardized metadata for human pathogen/vector genomic sequences," PloS One. vol 9 no 6, pp. e99979, June 2014.

[4] 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," Nat Biotechnol. vol. 25, pp. 1251-5, November 2007.

[5] M. Horridge, T. Tudorache, C. Nuylas, J. Vendetti, N. F. Noy, and M. A. Musen.. "WebProtégé: a collaborative Web-based platform for editing biomedical ontologies," Bioinformatics. vol. 30, pp. 2384-5, August 2014.

[6] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, et al., "The Ontology for Biomedical Invetigastions," PLoS One. vol 11 no. 4, pp. e0154556, April 2016.

[7] The Phenotype And Trait Ontology (PATO) [online]. Available: https://github.com/pato-ontology/pato/

[8] The Ontology for General Medical Sciences (OGMS) [online]. Available: https://github.com/OGMS/ogms/

[9] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, and S. E. Lewis, "The environment ontology: contextualising biological and biomedical entities," J. Biomed. Sem. vol. 4, pp. 43, December 2013.

[10] W.A. KIbbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, et al., "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," Nucleic Acids Res. vol. 43, pp. D1071-8, January 2015.

[11] J. Hanna, E. Joseph, M. Brochhausen, and W. R. Hogan, "Building a drug ontology based on RxNorm and other sources," J. Biomed. Sem. vol. 4, pp. 44 , December 2013.

[12] L. G. Cowell and B. Smith, "Infectious disease ontology," in Infectious disease informatics, Springer New York, 2010, pp. 373-395.

[13] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease," Am. J. Hum. Genet. vol. 83, pp. 610-5, November 2008.

[14] The Information Artifact Ontology (IAO) [Online]. Available: https://github.com/information-artifact-ontology/IAO/

[15] M. Brochhausen, J. Zheng, D. Birtwell, H. Williams, A. M. Masci, H. J. Ellis, et al., "OBIB – a novel ontology for biobanking," J. Biomed. Sem. vol. 7, pp. 23, May 2016.

[16] The Symptom Ontology (SYMP) [Online]. Available: http://symptomontologywiki.igs.umaryland.edu/mediawiki/index.php

[17] C. Jonquet, N. H. Shah, M. A. Musen, "The open biomedical annotator" Summit on Translat Bioinforma. vol. 2009, pp. 56-60, March 2009.

[18] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, et al., "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access

and use ontologies in software applications," Nucleic Acids Res. vol. 39, pp. W541-5, July 2011.

[19] Protégé [Online]: Available: http://protege.stanford.edu

[20] R. Arp, B. Smith, and A. D. Spear, "Building ontologies with Basic Formal Ontology," The MIT Press, 2015.