Semantic Modeling for Accelerated Immune Epitope Database (IEDB) Biocuration

Gully A Burns
Information Sciences Institute
Marina del Rey, California
burns@isi.edu

Randi Vita
La Jolla Institute for Allergy and Immunology
La Jolla, California
rvita@lji.org

James Overton
Toronto, Ontario, Canada
james@overton.ca

Ward Fleri
La Jolla Institute for Allergy and Immunology
La Jolla, California
wfleri@lji.org

Bjoern Peters
La Jolla Institute for Allergy and Immunology
La Jolla, California
bpeters@lji.org

ABSTRACT

The Immune Epitope Database (IEDB) indexes and organizes published information pertaining to the molecular targets of adaptive immune responses to support epitope discovery efforts. The IEDB is an exemplary system with a well-designed repository, a commercial-grade user interface and a large user community. It is expressly 'built-for-purpose', with a specialized Entity-Relation (ER) schema designed specifically to describe experimental findings (in this case, outcomes from assays relevant to immune epitope studies). Like many biomedical databases, this use of a specialized ER model impacts the process of indexing and organizing available scientific information. Biocuration staff and end-users must be trained specifically in the details of the representation to populate and use the system. The extent of system interoperability is generally limited to the use of standard terminology. We apply a knowledge engineering modeling methodology called "Knowledge Engineering from Experiment Design" (KEfED) that uses a workflow-like construct to model studies that had been curated into the IEDB. This methodology generates a semantic model for experimental data from dependency relations between experimental variables based on an experiment's protocol. We also applied the Karma mapping system to build a linked data representation of IEDB content across the whole database as a potential methodology for exporting IEDB content to a linked data format. This work demonstrates the feasibility of using KEfED modeling to represent previously-curated data in existing systems and then mapping that existing dataset to a linked data model. This may offer a graceful method for the evolution of existing, well-established databases.

CCS CONCEPTS

• **Information systems** → *Network data models*; *Information integration*;

KEYWORDS

Immune Epitopes, Knowledge Engineering, Biocuration

K-CAP2017 Workshops and Tutorials Proceedings,

ACM Reference Format:
Gully A Burns, Randi Vita, James Overton, Ward Fleri, and Bjoern Peters. 2018. Semantic Modeling for Accelerated Immune Epitope Database (IEDB) Biocuration. In . ACM, New York, NY, USA, 6 pages. https://doi.org/

1 INTRODUCTION

As a scientific discipline, biomedicine is complex, multidisciplinary, continuously evolving and increasingly data-driven. This has lead to the development of literally thousands of biomedical databases across a large number of domains. In the field of molecular biology, the journal "Nucleic Acids Research" publishes an annual review of active molecular biology databases [6] with articles that describe each system and a managed online catalog of active systems[1]. This list includes several large-scale, international informatics projects. In particular, the 2017 review presents a "golden set" of 110 databases that have "consistently served as authoritative, comprehensive, and convenient data resources widely used by the community". In this paper, we describe preliminary modeling work within one of these database systems (the Immune Epitope Database, IEDB), to improve curation processes and permit a more standardized representation of experiment observations. Ultimately, we anticipate this work to permit graceful evolution of the systems' underlying data schema and biocuration model.

This work is a principled approach to data modeling using "metadata propagation" through experimental workflows describing the physical processes in a laboratory experiment. Metadata propagation is a concept developed for e-Science workflow systems [7], but was repurposed as the driving principle of a flexible data modeling methodology for experimental data called "Knowledge Engineering from Experimental Design" (KEfED) [11].

The Ontology of Biomedical Investigations ("OBI") provides a mechanism for describing experimental protocols within the context of a well-defined upper ontology [2]). We previously developed an approach to modeling experimental variables [3] that we are currently integrating into OBI in order to apply KEfED modeling to data in the IEDB.

In this paper, we model a specific article that had been previously curated into the IEDB to act as a proof-of-concept of using the

[1] http://www.oxfordjournals.org/nar/database/a/

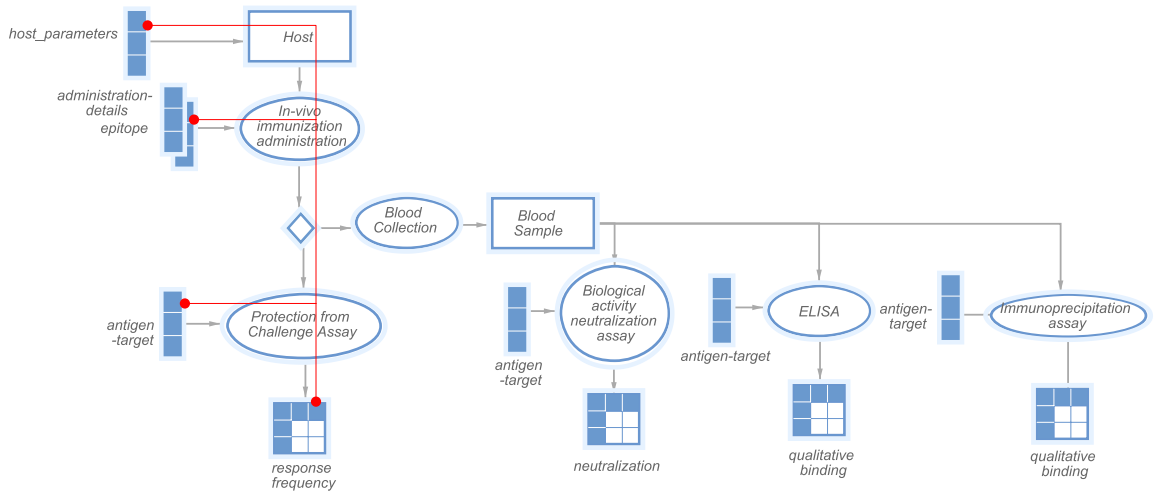


Figure 1: A KEfED model based on Richardson *et. al.* 1998. The red line shows dependency relations between measurements of 'response frequency' from a 'protection from challenge assay' and parameters set previously in the protocol.

KEfED methodology. We also describe use of the Karma data integration tool[2] as a way of automatically populating a KEfED-driven linked data representation.

2 METHODS

KEfED modeling work was performed with the "kefed.io" toolset[3]. We downloaded the latest versions of IEDB[4] and evaluated the IEDB schema and content in consultation with curation staff. We referenced IEDB's use of OBI ontology terms for assays in this modeling effort, whilst developing and proposing extensions to OBI for data item and value specification classes in order to provide adequate coverage for appropriate variables and associated values within the KEfED models under development.

An intermediate target for this modeling work was to provide a KEfED-based design pattern that could be used to convert IEDB data to linked data using ISI's Karma information integration tool [8]. We queried the B-Cell table from the IEDB for data from "protection from challenge" assays and then mapped columns from that data set onto the values of variables generated from our manually-curated KEfED model for the same class of experiments. This provided a viable procedure to migrate existing data from IEDB to linked data generated under a KEfED-based model. All modeling work was performed by hand and this effort was executed as a proof of concept for subsequent development.

3 RESULTS

Under development since 2004, the IEDB has undergone three large scale iterations to provide coverage of >95% of the relevant experimental biomedical literature. At present (November 30th 2017), it lists records from 18,902 journal articles focused on infectious diseases, allergy, autoimmunity and transplantation. HIV-derived and cancer epitopes are considered out of scope for this system as they are managed elsewhere[5]. Our goal in this work was to explore the feasibility of developing semantic models within the KEfED modeling formalism that could *reconstruct* the logic of the data that the IEDB currently contains.

As an advanced scientific database, the IEDB is based on complex, domain-specific knowledge. A key structural design concept that permits the capture of data from a wide number of different types of immunological experiments is the IEDB's use of well-defined *assay types*[6]. These are experimental processes that generate specific types of measurements with well-defined meanings that serve as the basic building blocks of immunological studies. The IEDB's set of assay types is also documented as classes in OBI[7] providing a well-defined base vocabulary to build upon.

3.1 Richardson *et al.* 1998: A Worked Example

We focus on one study in particular: Richard *et al.* 1998 [10]. A KEfED model that illustrates the assays used in this study is shown in Figure 1. This study uses peptidergic epitopes derived from proteins found in envelope proteins of Feline Immunodeficiency Virus (FIV) as immunogens (*i.e.* to trigger an immune response). Animals that had been immunized with these epitopes were subsequently investigated with four assays that measured (A) whether the immunization process provides protection from the effects of a subsequent immune challenge; (B+C) the degree of antigen-antibody binding occurring after the immunization step, measured either by (B) immunoprecipitation or (C) an ELISA step; finally (D) whether antibodies generated from experimental subjects were themselves capable of neutralizing FIV in a test environment.

Structurally, when viewed at this level, the design is simple. The host animal is immunized and a blood sample is drawn and processed with biological activity and binding assays. In addition, the same immunized host is subjected to a "protection from challenge

[2]http://karma.isi.edu
[3]https://github.com/SciKnowEngine/kefed.io
[4]available from http://www.iedb.org/database_export_v3.php
[5]https://www.hiv.lanl.gov/content/immunology/
[6]https://help.iedb.org/hc/en-us/articles/114094147271-IEDB-Assay-Types-IEDB-3-0-
[7]http://obi-ontology.org/

assay" to assess how well the immunization process protects the animal from FIV infection. The primary technical challenge of this work arises from the definition of variables that are relevant to the IEDB curation process.

In Figure 1, we provided variables with simple names ("host-parameters", "administration-details", etc.) to denote composite data structures that mirrored the relevant substructure of data pertaining to IEDB-relevant data. An example of this substructure is shown in Table 1 for the parameter "epitope" denoted in Figure 1 as an input to the "in-vitro immunization administration" process. Note that this substructure exactly corresponds to the data provided by the IEDB in their assay pages (for example: http://www.iedb.org/assay/1508651). By capturing the data structure used in IEDB directly into parameters, we are able to match the KEfED modeling approach precisely to the data described in IEDB. This effort is intended to supplement existing biocuration efforts at IEDB [5] and so will be evaluated within their framework for quality control.

Table 1: Substructure of 'epitope' variable

Sub-Parameter	Type	Example Value
epitope-type	category	linear peptide
linear-sequence	string	RAISSWKQRNRWEWRPD
start-position	integer	387
end-position	integer	403
source-name	string	Envelope glycoprotein gp150
source-accession	URI	ncbi-protein:Q05312.1
source-organism	URI	ncbi-taxon:45409
source-org-name	string	FIV (isolate wo)

3.2 Representing KEfED Models using OBI elements

Within the scope of established ontologies describing experimental methodology in the biomedical community, OBI is likely the most mature and well-supported [2]. Despite being linked to and incorporated in several other projects within the community (see https://bioportal.bioontology.org/ontologies/OBI), there is no single recommended methodology of how to use OBI terms to describe an experimental workflow. We therefore developed a schema for OBI-like elements that could capture the crucial elements of a KEfED model. Figure 2 shows this schema formatted as a UML2.0 class diagram. The purpose of this schema is to provide a framework for developing KEfED models that could act as templates made up of OBI-compatible terminology.

Consistent with OBI's extension of the Basic Formal Ontology ('BFO') [1], this schema extends the `Continuant` class to define `Material Entity` and `Data Item` classes. These elements are entities within the workflow that have continued existence over time. We also define `Planned_Process` elements that map directly to OBI's`Material Processing`, `Assay`, and `Data_Processing` classes. These elements denote key KEfED elements to describe the workflow. Less-well defined is the way in which the values of each data item is defined. Here, consistent with ongoing discussions within the OBI community, we extend the `Value_Specification` class to support data of a variety of different types including ordinal, categorical and structured data. This corresponds to distinctions we previously defined in the 'Ontology of Experimental Variables and Value' (OoEVV) [3].

A key extension for KEfED is a representation of a data-driven context for each measurement made within the experiment. Within this design, this function is provided by the `Metadata_Context` class which simply links measurement and parameter values together via `parameterizes` and `has_context` properties.

3.3 Mapping data from 'Protection From Challenge' Experiments with Karma

The Karma system provides a methodology for rapidly mapping data sources to an OWL ontology acting as a schema for linked data [8]. We executed a native SQL query over several IEDB tables (`article`, `bcell`, `object`, and `assay_type`) to retrieve data pertaining to "Protection from Challenge" assays across the whole database. This query retrieved 2,000 rows of data that specified "in vivo assay measuring B cell epitope specific protection from challenge" (term URI: http://purl.obolibrary.org/obo/OBI_0001710) or its subtypes as their assay type. We extended OBI with OWL classes corresponding to missing elements shown in Figure 2 and constructed a Karma model that mapped the extracted data to this extended KEfED/OBI ontology. Figure 3 provides a screenshot of a subset of the Karma model showing a portion of the mapping. The Karma interface uses the term URI as the primary label on the model display but will also show the label of the term if the user mouses over the term's node in the user interface. Modeling work was performed on a 2.5 GHz Intel Core i7 Macbook Pro with 16 GB RAM.

3.4 The Granularity of Processes: Expanding the "Protection from Challenge" Assay

Finally, we consider that descriptions of experimental processes have an inherent granularity based on the degree of detail that is required. We highlight this question by considering the 'Protection from Challenge' assay shown in Figure 1. A detailed reading of the paper, reveals that the assay as described in IEDB is actually made up of a number of individual steps that included (a) an immune challenge, (b) extraction of tissue and blood from the host, (c) RNA extraction and (d) subsequent competitive PCR to establish measures of viremia. This is significant since these intermediate steps involve data sets that form the main evidence presented by the paper's authors (measures of viremia in blood and spleen) that themselves must be evaluated to generate the final data item to be curated into the IEDB: "response frequency".

This is illustrated in detail in Figure 4 showing how, in this paper, the assay has quite a complex substructure at this intermediate level. It is also worth noting that many of these processes would themselves have detailed substructure that may of significance to a researcher maintaining their own laboratory-based record of experimental work with a very level of detail. We model this structure by permitting `Planned_Process` class instances to have `has_part` relations with other `Planned_Process` instances. This would permit multiple levels of sub-processes to be described in modeling of experimental protocols.

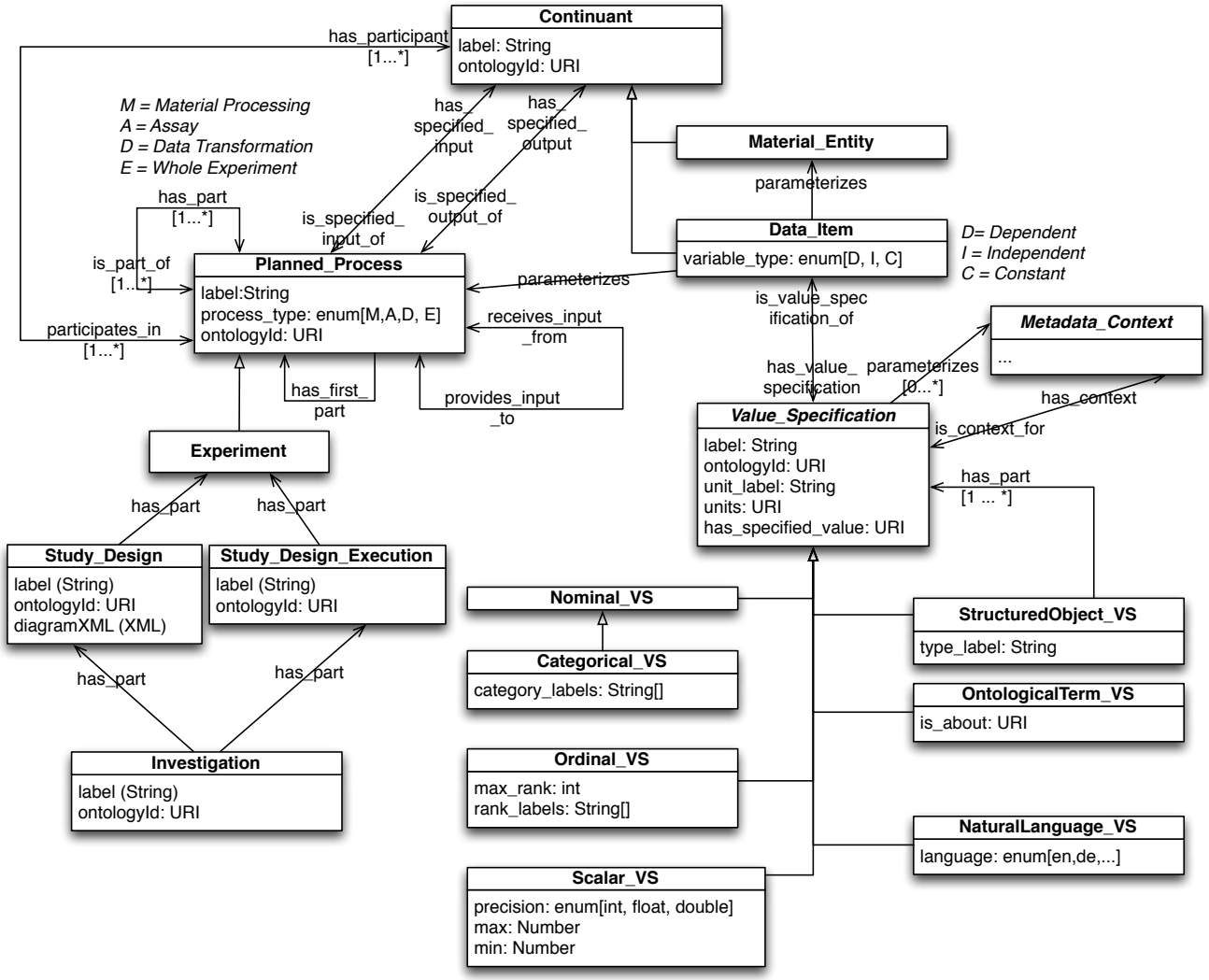


Figure 2: A data schema for representing KEfED models and data.

4 RELATED WORK

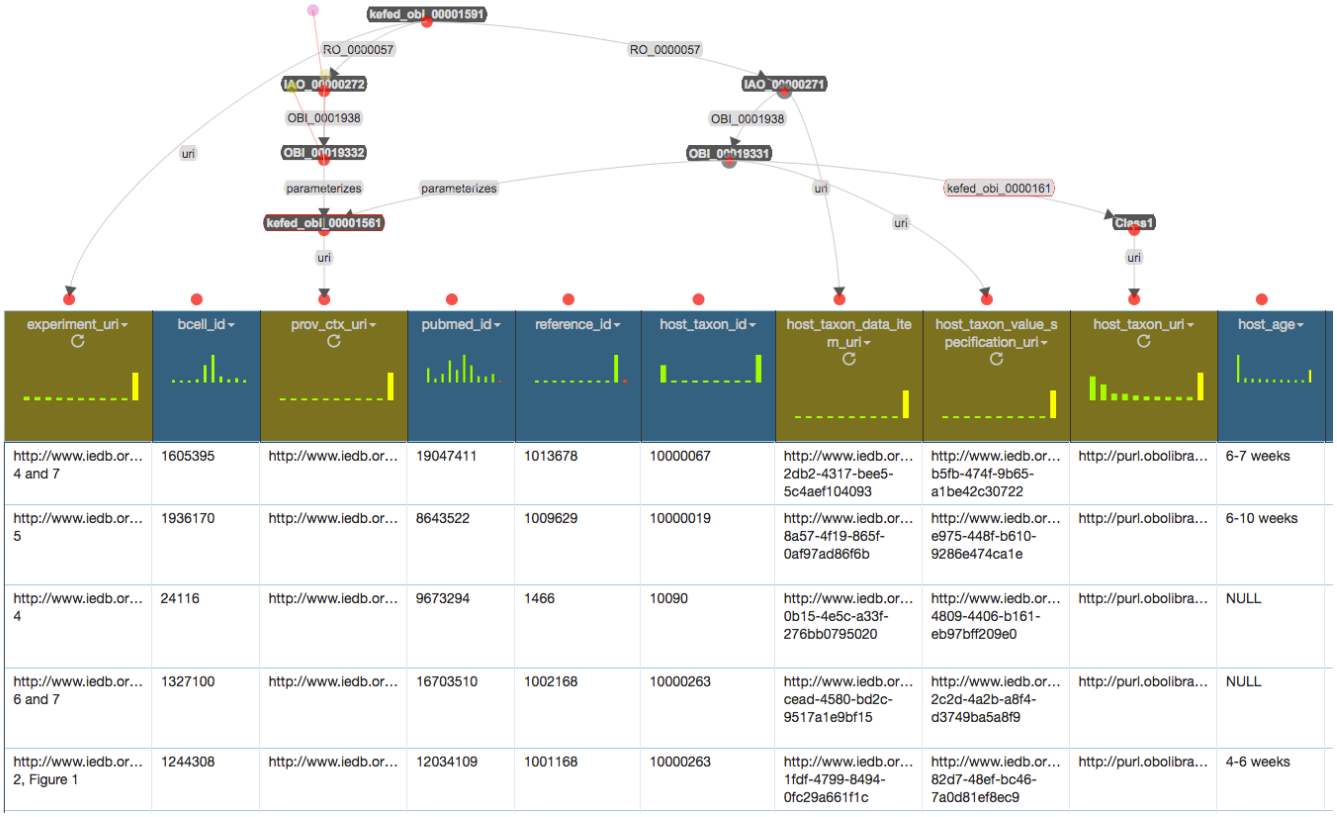
The IEDB uses OBI to support query formation within its user interface [14]. There are other ontological representations of protocols that complement this work. The Bioassay ontology (BAO) provides a representation of chemical biology screening assays [13]. The Evidence Ontology (ECO) provides a high-level ontological representation of different types of evidence used by biologists to draw conclusions that ties closely to OBI [4]. The latest version of the Experiment Action Ontology (EXACT2) incorporates OBI and constructs and focuses on representing the most granular actions (incubate, heat, *etc*) [12]. SMART Protocols provides a methodology originally derived from models of provenance[8]. STAR Methods is a publisher-initiated attempt to standardize terminology describing methodological resources used in biology [9]. None of these representations deal with the structure of claims at the level of the low-level variables that form the core of the KEfED representation.

5 DISCUSSION

This paper describes a simple proof-of-concept analysis of using the KEfED modeling approach as a possible methodology for improving the accuracy and speed of biocuration for an established biomedical database. Though far from definitive, this early work provides support to the notion that KEfED methods may effectively provide a general method of capturing scientific knowledge from published experimental studies *at a level of granularity that matches established databases such as the IEDB.*

A possible area of difficulty in applying KEfED to experimental findings in the literature is that there are typically a wide variety of experiments performed in any given subdomain. A database such

[8]https://smartprotocols.github.io/



experiment_uri	bcell_id	prov_ctx_uri	pubmed_id	reference_id	host_taxon_id	host_taxon_data_item_uri	host_taxon_value_specification_uri	host_taxon_uri	host_age
http://www.iedb.or… 4 and 7	1605395	http://www.iedb.or…	19047411	1013678	10000067	http://www.iedb.or… 2db2-4317-bee5-5c4aef104093	http://www.iedb.or… b5fb-474f-9b65-a1be42c30722	http://purl.obolibra…	6-7 weeks
http://www.iedb.or… 5	1936170	http://www.iedb.or…	8643522	1009629	10000019	http://www.iedb.or… 8a57-4f19-865f-0af97ad86f6b	http://www.iedb.or… e975-448f-b610-9286e474ca1e	http://purl.obolibra…	6-10 weeks
http://www.iedb.or… 4	24116	http://www.iedb.or…	9673294	1466	10090	http://www.iedb.or… 0b15-4e5c-a33f-276bb0795020	http://www.iedb.or… 4809-4406-b161-eb97bff209e0	http://purl.obolibra…	NULL
http://www.iedb.or… 6 and 7	1327100	http://www.iedb.or…	16703510	1002168	10000263	http://www.iedb.or… cead-4580-bd2c-9517a1e9bf15	http://www.iedb.or… 2c2d-4a2b-a8f4-d3749ba5a8f9	http://purl.obolibra…	NULL
http://www.iedb.or… 2, Figure 1	1244308	http://www.iedb.or…	12034109	1001168	10000263	http://www.iedb.or… 1fdf-4799-8494-0fc29a661f1c	http://www.iedb.or… 82d7-48ef-bc46-7a0d81ef8ec9	http://purl.obolibra…	4-6 weeks

Figure 3: A screenshot taken from the Karma model showing the mapping between the OBI-derived KEfED ontology and data from the IEDB.

as the IEDB manages to circumvent this complexity through the expertise of trained biocurators, who map research findings into the database schema. By using KEfED, we match our representation *as closely as possible* to the experimental design reported in the paper by using a more flexible data structure as a target of biocuration. It is this semantic flexibility that provides a target that closely mirrors the existing schema of the IEDB to enable this methodology to be used in data curation. It remains an open question as to how much of the experimental idiosyncrasies of each study design should be modeled. The rule of thumb we apply is to use the minimum information needed to recreate the structured conclusions of the study. Interestingly, using KEfED to model existing biomedical databases' capabilities provides a possible evaluation methodology for future work. This would require a quantitative comparison of KEfED-based methods to existing database capabilities based on (A) schema verification / validation, (B) system performance, and (C) usability.

A key future aspect of this process of knowledge capture is to develop methods of machine reading capable of identifying and populating KEfED models automatically. This remains an important and difficult challenge problem.

ACKNOWLEDGMENTS

The authors would like to thank Sharayu Gandhi for her careful work on development of the kefed.io Javascript interface. The work was performed under subcontract directly funded by the La Jolla Institute For Allergy and Immunology.

REFERENCES

[1] Robert Arp, Barry Smith, and Andrew D. Spear. 2015. *Building Ontologies with Basic Formal Ontology.* The MIT Press.

[2] Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, Melanie Courtot, Dirk Derom, Michel Dumontier, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Alejandra Gonzalez-Beltran, Melissa A. Haendel, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Mark Jensen, Yu Lin, Allyson L. Lister, Phillip Lord, James Malone, Elisabetta Manduchi, Monnie McGee, Norman Morrison, James A. Overton, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Daniel Schober, Barry Smith, Larisa N. Soldatova, Christian J. Jr Stoeckert, Chris F. Taylor, Carlo Torniai, Jessica A. Turner, Randi Vita, Patricia L. Whetzel, and Jie Zheng. 2016. The Ontology for Biomedical Investigations. *PloS one* 11, 4 (2016), e0154556. https://doi.org/10.1371/journal.pone.0154556

[3] Gully A P C Burns and Jessica A Turner. 2013. Modeling functional Magnetic Resonance Imaging (fMRI) experimental variables in the Ontology of Experimental Variables and Values (OoEVV). *Neuroimage* (May 2013). https://doi.org/10.1016/j.neuroimage.2013.05.024

[4] Marcus C. Chibucos, Christopher J. Mungall, Rama Balakrishnan, Karen R. Christie, Rachael P. Huntley, Owen White, Judith A. Blake, Suzanna E. Lewis, and Michelle Giglio. 2014. Standardized description of scientific evidence using

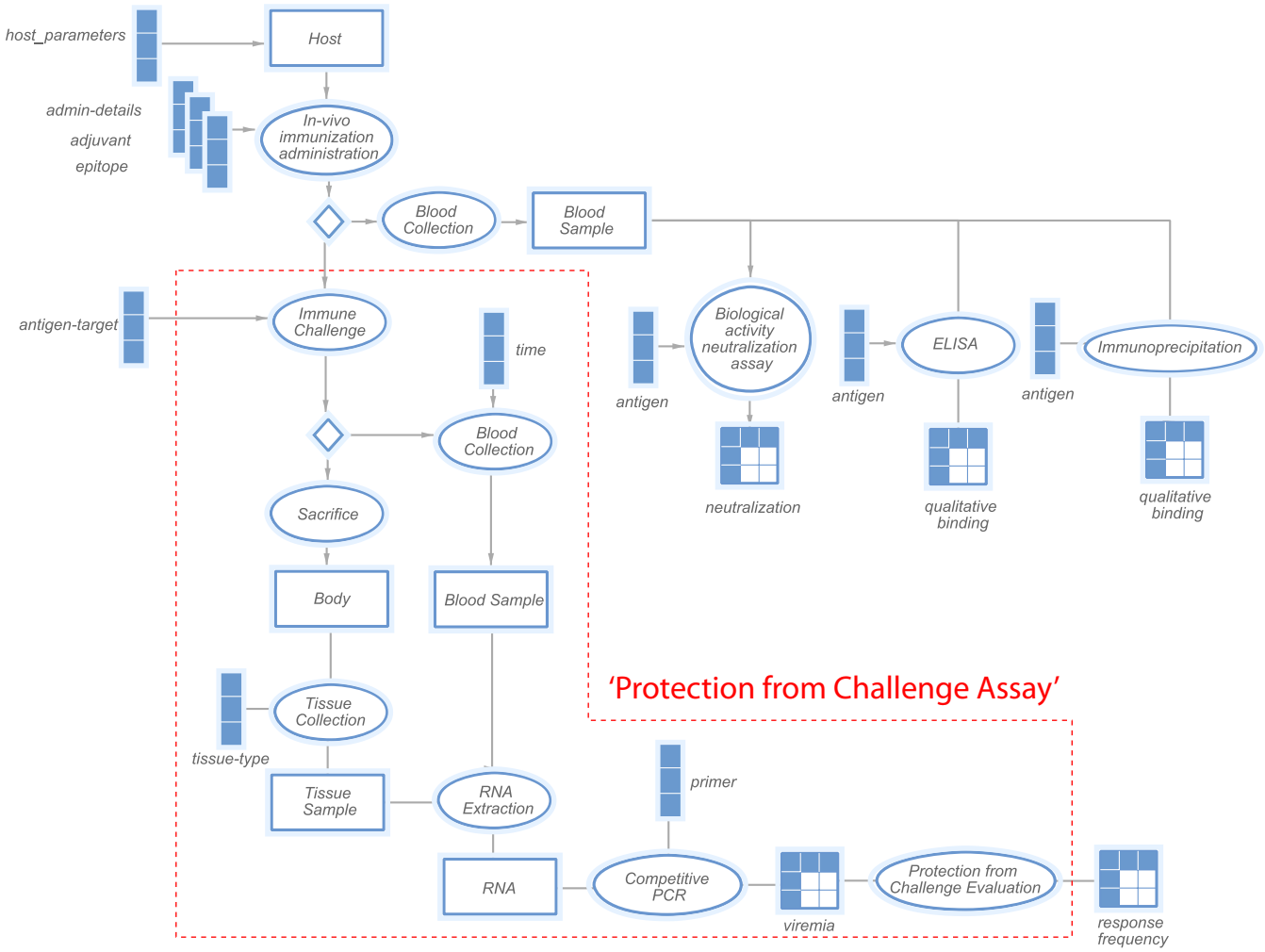


Figure 4: An expanded KEfED model that shows the internal substructure of the "Protection from Challenge Assay" process node for this experiment.

the Evidence Ontology (ECO). *Database : the journal of biological databases and curation* 2014 (2014). https://doi.org/10.1093/database/bau075

[5] Ward Fleri, Kerrie Vaughan, Nima Salimi, Randi Vita, Bjoern Peters, and Alessandro Sette. 2017. The Immune Epitope Database: How Data Are Entered and Retrieved. *Journal of immunology research* 2017 (2017), 5974574. https://doi.org/10.1155/2017/5974574

[6] Michael Y. Galperin, Xose M. Fernandez-Suarez, and Daniel J. Rigden. 2017. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. *Nucleic acids research* (Jan. 2017). https://doi.org/10.1093/nar/gkx021

[7] Yolanda Gil. 2014. Intelligent Workflow Systems and Provenance-Aware Software. In *Proceedings of the Seventh International Congress on Environmental Modeling and Software*. San Diego, CA. http://www.isi.edu/~gil/papers/gil-iemss14.pdf

[8] Craig A. Knoblock, Pedro Szekely, Jose Luis Ambite, and Aman Goel, Shubham Gupta, Kristina Lerman, Maria Muslea, Mohsen Taheriyan, and Parag Mallick. 2012. Semi-Automatically Mapping Structured Sources into the Semantic Web. In *Proceedings of the Extended Semantic Web Conference*. Crete, Greece.

[9] Emilie Marcus. 2016. A STAR Is Born. *Cell* 166, 5 (Aug. 2016), 1059–1060. https://doi.org/10.1016/j.cell.2016.08.021

[10] J. Richardson, A. Moraillon, F. Crespeau, S. Baud, P. Sonigo, and G. Pancino. 1998. Delayed infection after immunization with a peptide from the transmembrane glycoprotein of the feline immunodeficiency virus. *Journal of virology* 72, 3 (March 1998), 2406–2415.

[11] Thomas Russ, Cartic Ramakrishnan, Eduard Hovy, Mihail Bota, and Gully Burns. 2011. Knowledge Engineering Tools for Reasoning with Scientific Observations and Interpretations: a Neural Connectivity Use Case. *BMC Bioinformatics* 12, 1 (2011), 351. https://doi.org/10.1186/1471-2105-12-351

[12] Larisa N. Soldatova, Daniel Nadis, Ross D. King, Piyali S. Basu, Emma Haddi, Veronique Baumle, Nigel J. Saunders, Wolfgang Marwan, and Brian B. Rudkin. 2014. EXACT2: the semantics of biomedical protocols. *BMC bioinformatics* 15 Suppl 14 (2014), S5. https://doi.org/10.1186/1471-2105-15-S14-S5

[13] Ubbo Visser, Saminda Abeyruwan, Uma Vempati, Robin P. Smith, Vance Lemmon, and Stephan C. Schurer. 2011. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics* 12 (2011), 257. https://doi.org/10.1186/1471-2105-12-257

[14] Randi Vita, James A. Overton, Jason A. Greenbaum, Alessandro Sette, and Bjoern Peters. 2013. Query enhancement through the practical application of ontology: the IEDB and OBI. *Journal of biomedical semantics* 4 Suppl 1 (April 2013), S6. https://doi.org/10.1186/2041-1480-4-S1-S6