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

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

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

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.

KEYWORDS

Immune Epitopes, Knowledge Engineering, Biocuration

CCS CONCEPTS

• Information systems → Network data models; Information integration;

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SciKnow’17, December 2017, Austin, Texas, USA
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-x-yyyy-zzzz-y/YY/MM... $15.00
https://doi.org/10.1145/nnnnnn.nnnnnn

1 INTRODUCTION

As a scientific discipline, biomedicine is complex, multidisciplinary, continuously evolving and increasingly data-driven. This has led 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 systems1. 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
current 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 KEfED methodology. We also describe use of the Karma data integration tool as a way of automatically populating a KEfED-driven linked data representation.

2 METHODS

KEfED modeling work was performed with the “kefed.io” toolset. We downloaded the latest versions of IEDB 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. 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. 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. 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 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 a subsequent immune challenge; (B+C) the degree of antigen-antibody binding occurring after the immunization step, measured either by

---

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.
Within the scope of established ontologies describing experimental workflows, arises from the definition of variables that are relevant to the IEDB biocuration process. 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 OBI terms, 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 and will be evaluated within their framework for quality control.

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

### 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 UML 2.0 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.

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 provenance8. 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

---

8https://smartprotocols.github.io/
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 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
Figure 4: An expanded KEfED model that shows the internal substructure of the “Protection from Challenge Assay” process node for this experiment.