

## ADDENDUM

# Preliminary Meta-Analyses of Experimental Design with Examples from HIV Vaccine Protection Studies

Marcelo Tallis, Drashti Dave and Gully APC Burns

USC Information Sciences Institute  
4676 Admiralty Way, Suite 1001  
Marina del Rey, California 90292

### Abstract

Knowledge engineering from experimental design (KEfED) is a novel approach based on the dependency relationships that occur between the variables of a scientific study. We used this approach to curate the experimental designs of ten scientific papers from a well-established database of HIV vaccine trials in non-human primates. The KEfED models provide a characteristic, data-oriented signature for each measurement made in the study. We present preliminary analysis of these manually-curated, detailed representations using our own open-source curation tools and show the multi-variate statistical analyses on the resultant models of experimental design. The analyses produced a visualization of the similarities between studies and an account of the dependency relationships across studies. We describe our approach in the context of a knowledge engineering strategy based on creating large-scale domain-independent repositories of experimental observations.

In his classic work on the structure of scientific revolutions, Thomas Kuhn argues that the process of discovery is often triggered by crises sparked by anomalies between theory and experiment (Kuhn 1996). Developing a framework to capture these differences between interpretive predictions and observational data is therefore a key strategic goal for discovery informatics. Figure 1 shows our cyclic model for scientific reasoning, called ‘Cycles of Scientific Investigation or ‘CoSI’ (Russ et al. 2011; Helmer et al. 2011), echoing a similar view from other researchers (Clark and Kinoshita 2007; Soldatova and Rzhetsky 2011).

Within the CoSI model, the challenge of discovery informatics centers around *analyzing scientific knowledge to generate the key questions in a field that can then be investigated experimentally*. The analysis and investigation cycles typically requires the highest level of human expertise to execute and falls beyond the capabilities of current informatics systems to automate. We highlight this to provide a strategic goal and to contextualize our approach, described below.

The distinction between interpretations (expert, domain-specific knowledge required to construct theories, make predictions and plan experiments), and ‘observations’ (technically-defined details describing the protocols and the

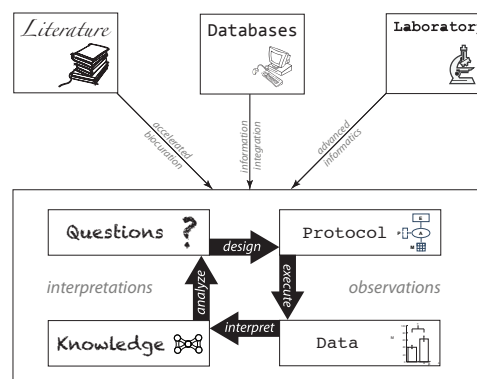


Figure 1: High-level design of the ‘CoSI model’.

data underlying experimental work) is usually overlooked within biomedical informatics applications. Whereas interpretations require expert knowledge, are more subjective and are highly domain-specific, we claim that observations require less expertise to define, are more objective (unless the experimental work described was poorly executed) and are not domain-specific. For example, neuroscientists, oncologists and marine biologists all use immunohistochemistry to show the distribution of proteins in mounted tissue. The interpretations of such data would vary depending on each context. Our aim to provide knowledge engineering tools that focus on widely reused observations in different interpretive contexts as informatics infrastructure.

The statistical assertions that form the basis of observations are grounded on the interactions of *variables*. These elements are used implicitly across almost all biomedical knowledge resources, including the tables, figures and text of research articles; structured column and table headings in databases; and some elements of community-derived ontologies and ‘minimum-information’ data models. They are ubiquitous, and yet there have been relatively few attempts to formulate systems that explicitly model and exploit experimental variables as fundamental building blocks.

Our work centers around one such attempt, called ‘Knowledge Engineering from Experimental Design’ (or

‘KEfED’). We focus on capturing the dependency relationships between variables forming the parameters, measurements and calculations of individual studies as the basis of a generic formalism (Russ et al. 2011). By looking closely at the structure of observational data within experimental protocols (represented using a relatively simple, lightweight approach), we are developing resources and approaches to biomedical informatics that are (a) comprehensible to biologists, (b) representative of the logical processes used by bench scientists to perform experiments and (c) generic enough to be applicable to fields that are not currently well-supported with extensive bioinformatics tools or systems.

## The LANL HIV Vaccine Trials Database

In 2009, roughly 0.5% of humanity was living with HIV / AIDS (assuming a global population of 6.8 billion at the time and an estimated number of people infected at 33.4 million; source: US Census and UNAIDS 2009). A safe, effective vaccine for HIV could save tens of millions of lives and non-human primates provide the best animal model for this disease. A prominent system cataloging HIV-related vaccination protection studies in primates is the HIV Trial Database of the Los Alamos National Laboratories (LANL) (Foley et al. 2007). The database contains high-quality, hand-curated accounts of HIV vaccine protection studies in non-human primates that provides us with an excellent use-case for the KEfED approach. Vaccine protection studies are comprised of a wide variety of interventions, assays, entities, parameters and measurements. The LANL system is an ideal vehicle that allows us to develop the KEfED approach as a general vehicle for structuring meta-analysis across studies.

## Systems and technology

Our goal is to provide functional software to the community for the construction, dissemination and usage of KEfED models within a laboratory’s day-to-day function. This includes the development of an ‘ontology design pattern’ called ‘the ontology of experimental variables and values’ (OoEVV) to gather definitions of variables for use in KEfED models (Burns and Dave 2012). The KEfED editing software in its current form provides a means for a biologist to draw a graphical model of the experimental protocol that tracks the dependency of measurements to be traced back to a set of parameters that form the overall ‘experimental context’ of a given measurement. An example of this dependency is shown in Figure 2 for measuring the cytotoxic T-lymphocyte (CTL) response in the post-immunization phase of a vaccination study (Belyakov et al. 2001). ‘CTL response’ (marked ‘6’ in Fig. 2) is a measurement that depends on the parameters marked ‘1’, ‘2’, ‘3’ and ‘4’.

Thus, the KEfED approach provides us with the means of analyzing the underlying structure of experiments in a principled way. The dependency signatures of measurements across multiple experiments vary as the organization and values of parameters are chosen to address different research questions. At this early stage, we use simple exploratory multivariate statistical approaches such as multidimensional scaling (Cox and Cox 1995) to examine the patterns within a

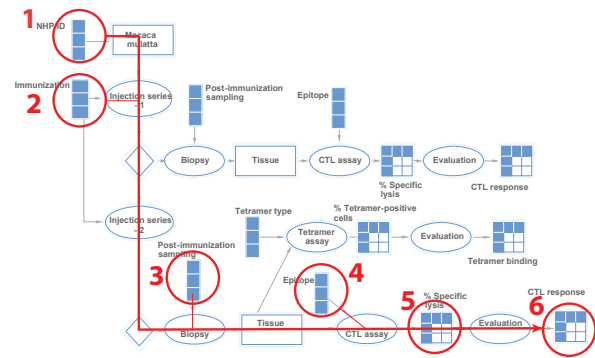


Figure 2: Dependency relations between variables in (Belyakov et al. 2001).

collection of ten detailed KEfED models of individual studies. We will describe the software development used to construct our systems; the curation and modeling work needed to construct the models; and the analysis approaches used for the results reported here. We will describe our findings and conclude with a discussion of related and future work.

## Methods

### KEfED software development

Our current and future development work are constructed on the View Primitive Data Model framework (VPDMf), a system that uses an object-oriented data model with graph-based ‘views’ that traverse an application-specific data model. VPDMf provides an encapsulation mechanism for querying and manipulating data across classes (Burns et al. 2003). In a similar way to the ‘Ruby on Rails’ agile approach to developing web-applications, the VPDMf acts as a scaffolding framework to generate source code (within multiple implementation environments in our case) that supports interoperability between different components by virtue of being created from the same underlying design. It allows us to use a single schema as the definition of a MySQL database, Java classes, an OWL ontology (and other components for other applications). The underlying UML design for the KEfED system is shown in Figure 3.

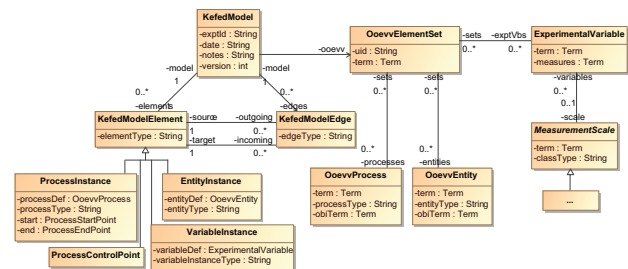


Figure 3: UML representation of the KEfED and OoEVV models.

Each individual KEfED model element is simply a typed node (denoting processes, entities, control points and vari-

ables in a specific experimental protocol) in a graph structure that links to definitions from the OoEVV system. Thus, each **OoevvElementSet** provides a vocabulary of elements that each individual **KefedModel** joins together in a particular configuration. Note that the current distribution of the full KEfED editing tool was built with a slightly different (but fully compatible) data model and we are in the process of transitioning that tool to the design in Figure 3.

## Manual curation of vaccine protection studies from the LANL database

We performed manual curation of KEfED models directly from research articles already curated into the LANL database. For each study from the LANL database, we have a means to check if the conclusions drawn by the domain experts at LANL can be derived from our emergent models. For our curation, we first identified the basic constants of the study (such as the species, vaccine and challenge), identified the immunization regimen and noted the techniques used along with their associated variables. Next, we sketched a logical flowchart of the protocol and then transferred this to the KEfED editor software based on all elements of the protocol. Finally, we instantiated the constants and parameters with data from the paper. For each paper, we have constructed the structure of the experimental design and instantiated the values of parameters, but not the data resulting from each experiment.

**KEfED modeling makes reading papers easier.** The time taken for the curation of each paper was roughly equivalent to that required to read it in depth (assuming the curator is trained in the use of our modeling tool). This process of model construction provided a structure that immediately supported comprehension of the data itself (see Figure 4).

In its 2007 report (Foley et al. 2007), the LANL vaccine trials database curated summary statements to represent the main findings of studies as natural language assertions. The first statement is as follows: “*Mucosal SIV specific CTL can be induced by intrarectal immunization of macaques with synthetic-peptide vaccine coupled with LT(R192G) adjuvant*” (case NHP.1, Finding 1). The assertion or main finding was an interpretation of the data shown in Figure 4. When examining the design of the complex data required to support the assertion, it is apparent that the values of the parameters and measurements traced in the KEfED model from Figure 2 directly correspond to the data reported in the paper. Thus, we are able to represent each data-point individually within the KEfED model (the numbered axis labels, 1-5, directly correspond to variable definitions shown in Fig. 2). We curated ten papers to provide the corpus for our subsequent analysis: (Belyakov et al. 2001; Patterson et al. 2001; Cafaro et al. 2001; Fuller et al. 2002; Muthumani et al. 2003; Buge et al. 2003; Lun et al. 2004; Rosati et al. 2005; Belyakov et al. 2006; Gomez-Roman et al. 2006).

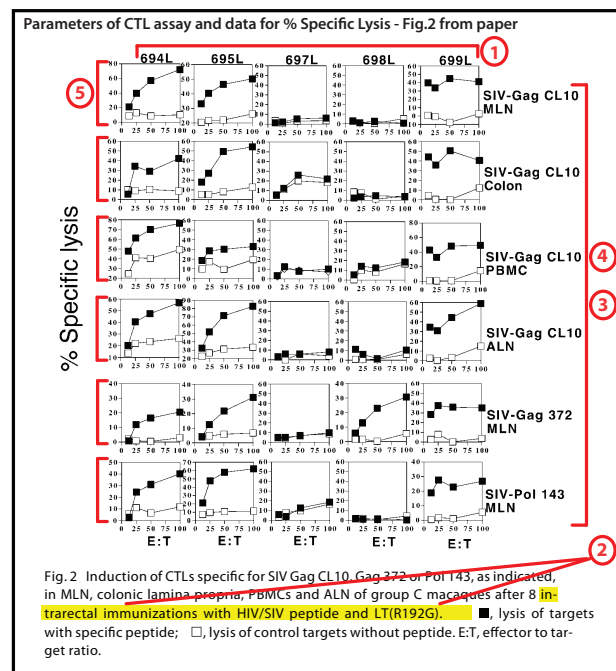


Figure 4: KEfED variables overlaid onto a data figure from (Belyakov et al. 2001) to showcase the utility of the KEfED methodology.

## Exploratory analysis of protocol structure across studies using nonmetric multidimensional scaling (NMDS)

We ran a set of summary queries over the KEfED database to count each occurrence of each type of element: processes, entities, parameters (both with and without data values) and measurements in each experiment. These lists of elements were simplified to provide binary features for each experiment that we used to calculate Jaccard similarity scores between studies (Cox and Cox 1995). We then used the parameter counts as features for similarity calculations between measurements. We then used the **isoMDS** command from the ‘R’ statistical analysis software to generate a two dimensional representation that treated these similarity scores as proximities within visualizations to illustrate relationships between experiments. We also reconstructed the full parameter-based signatures of each occurrence of each measurement and summed the number of experiments that a given measurement / parameter dependency was present to develop a statistical approach for examining this dependency.

## Results

### KEfED Model composition

The aggregated composition of OoEVV elements for the ten KEfED models were as follows: 8 entities (e.g., Blood sample, Tissue, etc.); 26 processes (e.g., Blood collection, ELISA, Sacrifice, etc); 49 measurements (e.g., Antibody concentration, Viral load, Antigenemia concentration, etc);

and 67 parameters (*e.g.*, ‘Post-challenge sampling.Time’, ‘CTL Prms.Stimulating epitope’ *etc*). Note that we grouped together parameters for each assay so that parameters that use the same measurement scale (*e.g.*, ‘time’, ‘antigen’, *etc*) in different contexts are considered to be different. We provide supplementary data of the full models within the KEfED modeling system as a web archive (\*.war) file.

### Multivariate analysis across studies

Figure 5 shows the two-dimensional NMDS configurations calculated from similarities derived from (a) all KEfED components or (b) parameters, as the features used to calculate proximities. Procrustes comparison between the two configurations revealed a root mean squared error of 0.047.

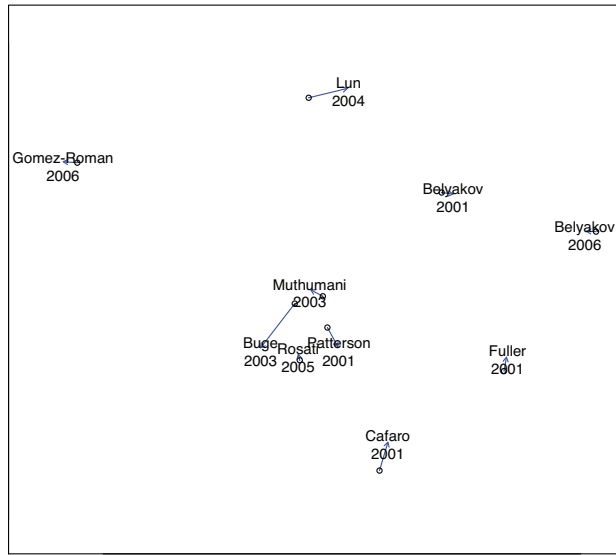


Figure 5: Two NMDS configurations of KEfED-derived models based with similarities based on overlap of all KEfED components (circles) and parameters (labeled arrowheads) related via Procrustes rotations. Labels are studies’ first author surname with the year of publication.

The layout of Figure 5 is consistent with some known aspects of the different designs of studies. There are two quite distinct outliers in this configuration (Gomez-Roman *et al.* 2006; Lun *et al.* 2004). Typically there are four stages when assays are used to evaluate different aspects of the immune response to viral challenge: (1) pre-immunization; (2) post-immunization, (3) post-challenge and (4) post-sacrifice. Firstly, the study by Gomez-Roman *et al.* was based on a different overall experimental setup. Here, the authors used animals from different studies (both infected and uninfected) to investigate the virulence of a specific type of vaccine vector. The Gomez-Roman study involved pre- and post-immunization but no post-challenge or post-sacrifice steps. Lun *et al.* perform a longitudinal study in which previously immunized monkeys are challenged, after which their post-challenged conditions are examined. The Lun study contrasts with the other studies that typically perform post-immunization, post-challenge and post-sacrifice evaluations.

All evaluations incorporate similar combinations of parameters across a variety of different measurements. Our interpretation of the grouping of the Buge, Rosati, Patterson and Muthumani studies is that they use the combination of approximately the same assays at various points in their experimental designs.

### Identifying consistent parameter signatures across studies

One of the goals of the current study is to identify consistent patterns of parameter signatures for measurements across studies as a means to assist with the development of an effective open-linked data model of scientific observations. The parameter signatures are generated as pathways through the KEfED model that always originate at the first entity. For vaccine protection studies, the first entity is invariably a nonhuman primate experimental subject (parameterized by ‘NHP ID’). Table 1 provides a summary of the dependency of each measurement on various parameters across models. Some measurements are made more than once at different stages in an individual experiment leading to an overall count greater than 10 (*i.e.*, ‘Viral load’ has a dependency count of 15 for ‘NHP ID’).

Table 1: Table of counts of measurements against indexing parameter set.

Measurements	Parameter sets	Challenge	CTL Prms	Cytotoxic Prms	Cytoviremia Prms	ELISA Prms	ELISPOT Prms	FACS Prms	Immunization	Neutralization Prms	NHP ID	Post-challenge sampling	Post-immunization sampling	Post-sacrifice sampling	Pre-immunization sampling	Proviral DNA Prms	Serotype Prms	Skin test Prms	Tetramer Prms	Viral load Prms	Virus isolation Prms
% Specific lysis		2	8																		
% Tetramer-positive cells																				2	
Antibody concentration		3			7																
Antibody response		4			12																
Antibody titers		1			4																
Antigenemia concentration		1			1																
Antigenemia presence		1			1																
Avidity ratio			1																		
Cell expansion		1																			
Cell numbers		1																			
CTL avidity			1																		
CTL response		2	8																		
Cytokine concentration					3																
Cytokine response					3																
Cytotoxic effect numbers				2																	
Cytoviremia numbers		1		1																	
Cytoviremia presence		1		1																	
ELISPOT numbers		1			4																
ELISPOT response		1			4																
Erythema																					
Erythema presence																					
Induration																					
Induration presence																					
Neutralizing antibody concentration		2																			
Neutralizing antibody index																					
Neutralizing antibody response		4																			
Neutralizing antibody titers		2																			
Optical density																					
Proliferative response		2																			
Proviral DNA copies		3																			
Proviral DNA presence		3																			
Serotype																					
Stimulation index		2																			
T cell counts		9																			
T cell numbers		9																			
Temperature																					
Temperature fluctuations																					
Tetramer binding																					
Viral DNA copies																					
Viral load		11																			
Viral presence		1																			
Viral RNA copies		11																			
Virus titers		1																			
Weight		1																			
Weight fluctuations																					

Table 1 shows a *preliminary* result: the dependency relationships between measurements and parameters for only a small number of experiments in a single field of study. Using experimental design as the guiding principle for building

data representation of data provides a powerful complete, ‘bottom-up’ approach for designing scientific databases, since each unique experiment may incorporate additional parameters of arbitrary complexity creating the situation that any database with a uniform design across multiple experiments would be (a) dealing with data from a single type of experiment or (b) simplifying the underlying data (as is the case with the LANL HIV Database). Our vision is to use this approach to provide semantic structure for KEfED-driven ‘nanopublications’ that represent precisely the meaning of the individual experiments that generated them (Groth, Gibson, and Velterop 2010).

## Discussion and Related work

The field of statistical natural language processing (NLP) was born when developers found that statistically-driven engineering approaches improved performance more than those provided by linguistic expertise (Jurafsky and Martin 2000). The core idea of our work is to follow the same principle in developing knowledge engineering technology for biomedical informatics by gathering statistics of experimental design across multiple studies.

We curated ten detailed KEfED models from primary research papers and used simple techniques to perform preliminary analyses. Attempting to represent observations faithfully ‘as they were originally intended’ highlights important technical difficulties underlying definitions of experimental variables: when observations are defined using different variables, both measuring the same quality, we use ‘measurement scales’ to distinguish between their values (Burns and Dave 2012). Attempts by the informatics community at standardization of variable definitions may restrict the ability of bench scientists to express their data and results. We cannot assume that our constructs are either valid or practically useful for scientists unless we survey them. A *descriptive* approach (rather than a *prescriptive* one), that delivers immediate practical value to scientists without requiring a large shift in data management practices is probably the most practical way to ensure the creation and adoption of standards. Incorporating evaluation metrics that directly address issues such as *usability* and *task performance* would support the uptake of standardized representations (Adelman and Riedel 1997).

In our work, we have attempted to assist with usability by providing a user-friendly tool to enable scientists to build semantic models of experimental protocols and observations within a conceptual framework that is technically comprehensible to bench scientists (Russ et al. 2011). We used a relatively simple semantic modeling approach to curate only ten experimental studies and started to investigate dependency relationships between variables using statistical approaches. We prioritize these dependency relations within our work, rather than the closely related issues of the provenance of data (Moreau et al. 2010) or the execution of workflows (Gil et al. 2007), since they define the structure of the statistical assertions that can be made from the data and are therefore crucial for subsequent analysis. Integrating the construction of these models seamlessly into the everyday workflow of scientists and informatics researchers is a high

priority and provides a continuation of earlier work developing literature-based knowledge management systems (Burns and Cheng 2006). We are also pursuing the use of KEfED as a framework for pre-publication data management (Jacobs et al. 2009).

The Ontology of Biomedical Investigation, or ‘OBI’ (Brinkman et al. 2010) is a community-driven ontology that seeks to provide terminological elements for use by any biomedical study. We matched our representation of protocol elements (processes, material- and information-entities) to the core definitions of OBI and intend the tools we construct to facilitate the development of this resource. Another important existing representation of measurements, units and factors is the Experimental Factor Ontology (EFO) (Malone et al. 2010). Similarly, the Vaccine Ontology (He and Xiang 2010) provides a set of practical terminology for vaccine protection studies (as well as an ontological representation of statistical analyses) providing a practical terminological repository for us to draw from and contribute to. At this early stage, we seek to standardize the terminology used in KEfED model elements with a relatively simple Ontology Design Pattern (Gangemi and Presutti 2009), called ‘the Ontology of Experimental Variables and Values’ (Oo-EVV) (Burns and Turner 2012). This system permits definitions of protocol elements, variables and scales to be curated and uploaded to the BioPortal ontology repository (see <http://bioportal.bioontology.org/ontologies/3006>).

The open provenance model (Moreau et al. 2010) is an effort to standardize provenance relationships supporting tracing where a dataset in a given study originates from. It does not capture the dependency relationships between variables that form the core contribution of the KEfED approach and interoperability with provenance tools is a priority for future work. Our current approach to tracing this dependency does not work for data transformations after the first act of direct measurement. We will address this by incorporating representations concerned with the input and output signatures of data processing steps in scientific workflows (Gil et al. 2007).

The concept of *nanopublications* provides a mechanism for publishing scientific findings as relatively succinct RDF fragments that make a scientific assertion (Groth, Gibson, and Velterop 2010). These computational elements should be typed according to the originating context of the assertion and will also include all necessary provenance information to attribute credit to the assertions authors. The KEfED model provides a framework for adding semantic structure to these assertions (Russ et al. 2011) and the mechanisms we describe here could be used to attempt to aggregate many such assertions and to look for patterns across them. Statistical NLP provides a rich toolset of analysis methods for probabilistic finite state automata (Jurafsky and Martin 2000) that may be directly applicable to the aggregation and analysis of KEfED models.

The presence of a statistically significant effect in experimental data is typically expressed as the upper limit of probability that the effect could be happening by chance (as a standard part of a significance test). The way that experimental studies probe the probabilistic dependencies between

study variables leads us to suggest that the KEfED approach could be used as a base for probabilistic graphical models to be used to reason over experiment data from biomedical experiments. Until now, Bayesian methods have mainly been applied to genetics and systems biology and there is interest in using them as a general methodology for scientific discovery (Kell 2012). Our formulation explicitly deals with variables using different (sometimes qualitative) measurement scales, and future work could investigate how KEfED might enable Bayesian Networks as a general approach for reasoning. Of particular interest are mathematical models of causality (Pearl 2000). In biology, experimentalists investigate causal relationships by manipulating their experimental paradigm to establish whether conditions composing a putative mechanism are *necessary* and/or *sufficient* for the resultant phenomenon. For example, a neuroscientist might block a neurotransmitter's action pharmacologically to see if a behavior disappears, demonstrating that the neurotransmitter's action is 'necessary' for the behavior. He may also use microinfusions of neurotransmitter to attempt to trigger the behavior to test the 'sufficient' condition. KEfED models could be used as the basis for studying scientific causality using counterfactuals (Pearl 2000).

Rather than attempting to drive the development of standardized database schema, minimum information data models and ontologies for specific domains by focusing on a human-derived knowledge representation, we propose an alternative approach: construct *detailed, accurate models of the structure and content of experimental observations* using the KEfED framework (or equivalent) to provide a domain-independent repository of experimental observations. This could then be queried to derive domain-specific representations based on scoped, statistically relevant semantic structures for interpretative models. Such a proposal is consistent with a biomedical semantic web approach. We also introduce the semantic distinction between observations and interpretations and provides structures for data. The challenge of this effort would be scalability since the only method we currently have to build such a system is manual curation. Existing RDF mashups can consist of billions of triples whereas we describe manual curation of only ten papers (Nolin et al. 2012). The comparison between our work and semantic web mashups misses several points: (A) the mashups are constructed from resources that had themselves been manually curated, (B) the vital provenance back to any supporting experimental data has almost certainly been lost in these resources and (C) there is no concrete guarantee that the knowledge included in the synthesis is even accurate.

Building informatics frameworks to assist discovery need to ensure that unlikely, fortunate, informative coincidences are both possible within the system and detectable once they do so. The scale of open-linked data in the existing semantic is large, but lacks the conceptual structure (based on experimental design) that scientists use to direct and drive their investigations. KEfED is one possible model to provide the conceptual backbone to support a large-scale 'liquid network' of scientific observations within which we could then construct automated systems for scientific discovery.

## Acknowledgement

This research is funded by NSF (#0849977) and by NIH (GM083871, MH079068 and RR025736). We thank Brian Foley and Bob Murnane for their support. We gratefully acknowledge the work of Cartic Ramakrishnan and Eduard H. Hovy in their support within discussions.

## References

- Adelman, L., and Riedel, S. L. 1997. *Handbook For Evaluating Knowledge-Based Systems*. Boston: Kluwer Academic Publishers.
- Belyakov, I. M.; Hel, Z.; Kelsall, B.; Kuznetsov, V. a.; Ahlers, J. D.; Nacsa, J.; Watkins, D. I.; Allen, T. M.; Sette, a.; Altman, J.; Woodward, R.; Markham, P. D.; Clements, J. D.; Franchini, G.; Strober, W.; and Berzofsky, J. a. 2001. Mucosal AIDS vaccine reduces disease and viral load in gut reservoir and blood after mucosal infection of macaques. *Nature medicine* 7(12):1320–6.
- Belyakov, I. M.; Kuznetsov, V. a.; Kelsall, B.; Klinman, D.; Moniuszko, M.; Lemon, M.; Markham, P. D.; Pal, R.; Clements, J. D.; Lewis, M. G.; Strober, W.; Franchini, G.; and Berzofsky, J. a. 2006. Impact of vaccine-induced mucosal high-avidity CD8+ CTLs in delay of AIDS viral dissemination from mucosa. *Blood* 107(8):3258–64.
- Brinkman, R.; Courtot, M.; Derom, D.; Fostel, J.; He, Y.; Lord, P.; Malone, J.; Parkinson, H.; Peters, B.; Rocca-Serra, P.; Ruttenberg, A.; Sansone, S.; Soldatova, L.; Stoeckert, C. J.; Turner, J.; and Zheng, J. 2010. Modeling biomedical experimental processes with obi. *J Biomed Semantics* 1(1):S7.
- Buge, S.; Ma, H.; Amara, R.; Wyatt, L.; Earl, P.; Villinger, F.; Montefiori, D.; Staprans, S.; Xu, Y.; Carter, E.; O'Neil, S.; and Herndon JG, Hill E, Moss B, Robinson HL, M. J. 2003. Gp120-alum boosting of a Gag-Pol-Env DNA/MVA AIDS vaccine: poorer control of a pathogenic viral challenge. *AIDS Res Hum Retroviruses* 19(10):891–900.
- Burns, G. A., and Cheng, W. C. 2006. Tools for knowledge acquisition within the neuroscholar system and their application to anatomical tract-tracing data. *J Biomed Discover Collab* 1(1):10. 1747-5333 (Electronic) Journal article.
- Burns, G. A., and Dave, D. 2012. A lightweight Ontology Design Pattern to curate and represent experimental variables from vaccine protection studies. In *To appear in Vaccine and Drug Ontology in the Study of Mechanism and Effect, ICBO 2012*.
- Burns, G. A., and Turner, J. A. 2012. Bottom-up curation of terminology for experimental variables: the ontology of experimental variables and values (OoEVV). In *Bio-Ontologies SIG, ISMB*.
- Burns, G. A.; Khan, A. M.; Ghandeharizadeh, S.; O'Neill, M. A.; and Chen, Y. S. 2003. Tools and approaches for the construction of knowledge models from the neuroscientific literature. *Neuroinformatics* 1(1):81–109. 1539-2791.
- Cafaro, A.; Titti, F.; Fracasso, C.; Maggiorella, M.; Baroncelli, S.; Caputo, A.; Goletti, D.; Borsetti, A.; Pace, M.; Fanales-Belasio, E.; Ridolfi, B.; Negri, D.; Sernicola, L.;

- Belli, R.; Corrias, F.; Macchia, I.; Leone, P.; Michelini, Z.; ten Haaf, P.; Butto, S.; Verani, P.; and Ensoli. 2001. Vaccination with DNA containing tat coding sequences and unmethylated CpG motifs protects cynomolgus monkeys upon infection with simian/human immunodeficiency virus (SHIV89.6P). *Vaccine* 19:2862–77.
- Clark, T., and Kinoshita, J. 2007. Alzforum and swan: the present and future of scientific web communities. *Brief Bioinform* 8(3):163–71.
- Cox, T. F., and Cox, M. A. A. 1995. *Multidimensional Scaling*. Monographs on Statistics and Applied Probability. Newcastle Upon Tyne: Chapman & Hall, 1 edition. Best Reference Book I've found on MDSThesis.
- Foley, B.; Hong-Geller, E.; Mokli, J.; Gupta, K.; Marthas, M.; Letvin, N.; and Korber, B. 2007. HIV/SIV Vaccine Trials Database. Technical report, Los Alamos National Laboratory, LA-UR 07-4296.
- Fuller, D.; Rajakumar, P.; Wilson, L.; Trichel, A.; Fuller, J.; Shipley, T.; Wu, M.; Weis, K.; Rinaldo, C.; Haynes, J.; and Murphey-Corb. 2002. Induction of mucosal protection against primary, heterologous simian immunodeficiency virus by a DNA vaccine. *J Virol* 76(7):3309–17.
- Gangemi, A., and Presutti, V. 2009. Ontology design patterns. In Staab, S., and Studer, R., eds., *Handbook of Ontologies*. Springer, 2nd edition.
- Gil, Y.; Deelman, E.; Ellisman, M.; Fahringer, T.; Fox, G.; Gannon, D.; Goble, C.; Livny, M.; Moreau, L.; and Myers, J. 2007. Examining the challenges of scientific workflows. *Computer* 40(12):24–32.
- Gomez-Roman, V.; Grimes, G.; Potti, G.; Peng, B.; Demberg, T.; Gravlin, L.; Treece, J.; Pal, R.; Lee, E.; Alvord, W.; and Markham PD, R.-G. 2006. Oral delivery of replication-competent adenovirus vectors is well tolerated by SIV- and SHIV-infected rhesus macaques. *Vaccine* 24(23):5064–72.
- Groth, P.; Gibson, A.; and Velterop, J. 2010. The anatomy of a nanopublication. *Information Services & Use* 30:51–56.
- He, Y., and Xiang, Z. 2010. Bioinformatics analysis of brucella vaccines and vaccine targets using violin. *Immunome Res* 6 Suppl 1:S5.
- Helmer, K.; Ambite, J.; Ames, J.; Ananthakrishnan, R.; Burns, G.; Chervenak, A.; Foster, I.; Liming, L.; Keator, D.; Macciardi, F.; Madduri, R.; Navarro, J.; Potkin, S.; Rosen, B.; Ruffins, S.; Schuler, R.; Turner, J.; Toga, A.; Williams, C.; and Kesselman, C. 2011. Enabling collaborative research using the Biomedical Informatics Research Network (BIRN). *J Am Med Inform Assoc*. 18(4):416.
- Jacobs, G.; Llovet, P.; Judson, I.; White, L.; Heimbuch, R.; Jeager, C.; and Burns, G. 2009. MILO: A general purpose data repository for disease foundations and individual laboratories. In *Society for Neuroscience Annual Meeting*.
- Jurafsky, D., and Martin, J. H. 2000. *Speech and Language Processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, Inc.
- Kell, D. B. 2012. Scientific discovery as a combinatorial optimisation problem: how best to navigate the landscape of possible experiments? *BioEssays: news and reviews in molecular, cellular and developmental biology* 34(3):236–244. PMID: 22252984.
- Kuhn, T. S. 1996. *The structure of scientific revolutions*. Chicago, London: University of Chicago Press, 3rd edition edition.
- Lun, W.; Takeda, A.; Nakamura, H.; Kano, M.; Mori, K.; Sata, T.; Nagai, Y.; and Matano, T. 2004. Loss of virus-specific CD4(+) T cells with increases in viral loads in the chronic phase after vaccine-based partial control of primary simian immunodeficiency virus replication in macaques. *J Gen Virol*. 85(7):1955–63.
- Malone, J.; Holloway, E.; Adamusiak, T.; Kapushesky, M.; Zheng, J.; Kolesnikov, N.; Zhukova, A.; Brazma, A.; and Parkinson, H. 2010. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics (Oxford, England)* 26(8):1112–8.
- Moreau, L.; Clifford, B.; Freire, J.; Futrelle, J.; Gil, Y.; Groth, P.; Kwasnikowska, N.; Miles, S.; Missier, P.; Myers, J.; Simmhan, Y. L.; Stephan, E.; and Van Den Bussche, J. 2010. The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems* to appear(August 2007):1–30.
- Muthumani, K.; Bagarazzi, M.; Conway, D.; Hwang, D.; Manson, K.; Ciccarelli, R.; Israel, Z.; Montefiori, D.; Ugen, K.; Miller, N.; and Kim J, Boyer J, W. D. 2003. A Gag-Pol/Env-Rev SIV239 DNA vaccine improves CD4 counts, and reduce viral loads after pathogenic intrarectal SIV(mac)251 challenge in Rhesus Macaques. *Vaccine* 21(7-8):629–37.
- Nolin, M.-A.; Dumontier, M.; Belleau, F.; and Corbeil, J. 2012. Building an HIV data mashup using Bio2RDF. *Briefings in bioinformatics* 13(1):98–106.
- Patterson, L. J.; Robey, F.; Muck, a.; Van Remoortere, K.; Aldrich, K.; Richardson, E.; Alvord, W. G.; Markham, P. D.; Cranage, M.; and Robert-Guroff, M. 2001. A conformational C4 peptide polymer vaccine coupled with live recombinant vector priming is immunogenic but does not protect against rectal SIV challenge. *AIDS research and human retroviruses* 17(9):837–49.
- Pearl, J. 2000. *Causality*. Cambridge: Cambridge University Press.
- Rosati, M.; von Gegerfelt, A.; Roth, P.; Alicea, C.; Valentin, A.; Robert-Guroff, M.; Venzon, D.; Montefiori, D.; Markham, P.; Felber, B.; and Pavlakis, G. 2005. DNA vaccines expressing different forms of simian immunodeficiency virus antigens decrease viremia upon SIVmac251 challenge. *J Virol* 79(13):8480–92.
- Russ, T.; Ramakrishnan, C.; Hovy, E.; Bota, M.; and Burns, G. 2011. Knowledge Engineering Tools for Reasoning with Scientific Observations and Interpretations: a Neural Connectivity Use Case. *BMC Bioinformatics* 12(1):351.
- Soldatova, L. N., and Rzhetsky, A. 2011. Representation of research hypotheses. *Journal of biomedical semantics* 2 Suppl 2(Suppl 2):S9.