Towards a Knowledge-intensive Framework for Efficient Vaccine Development

Leonard Petnga
Department of Industrial and Systems Engineering
University of Alabama in Huntsville
Huntsville, AL 35899, USA
Email: leonard.petnga@uah.edu

Surangi Jayawardena
Department of Chemistry
University of Alabama in Huntsville
Huntsville, AL 35899, USA
Email: surangi.jayawardena@uah.edu

Abstract—**Modern vaccine research & development efforts are complex, long, costly undertakes with high rate of failure. A major cause of inefficiency can be attributed to the mostly unstructured, unorganized, disconnected and diversity of knowledge spread across the Vaccine Development Life Cycle (VDLC) and honed by number of stakeholders with conflicting interests. State-of-the-art approaches have mostly fostered stove-piping knowledge and information within individual disciplines and separation of concerns, with little interest or appetite for cross-domain knowledge integration. In this research, we build on the ability of systems engineering to bridge the gaps between (and integrate) other disciplines to architect and develop a novel knowledge-intensive framework for efficient vaccine development. We formulate a model-based platform that accounts for the need for, (1) formalisms to support unambiguous and correct knowledge representation and reasoning across the VDLC, (2) capturing stochastic system biology behaviors and integrating with stakeholders' discrete decisions and, (3) models that are formal, reusable, customizable and can be assembled as needed for the purpose of the analysis at hand. Description logic-based formalisms and foundational domain theories support knowledge models of domains tightly coupled with Markov models of biological and chemical processes are the cornerstone of our framework. An example step-by-step implementation procedure illustrates the modularity, flexibility and configuration of the framework for tackling increasingly complex, cross-domain challenges across the VDLC. Vaccine preservation laboratory experiments are conducted to assess some prototype formulations and generate system biology models to be integrated with semantic models in the platform. Results are very encouraging but further work is needed in identifying and mapping all relevant biological system behaviors for the analysis under consideration and improving their characterization and integration in the framework.**

Keywords-Vaccine; Knowledge formalisms; Systems Engineering; Semantic; Ontology; Markov Chain; Multiple Regression Model.

I. INTRODUCTION

This work is concerned with the development and prototyping of a framework based on knowledge description formalisms and stochastic modeling of biological systems for improved efficiency across the Vaccine Development Lifecycle (VDLC). It stems from and, extends previous work on Knowledge-driven Vaccine Systems Engineering [1]. Scientific breakthroughs in biotechnology and genetic decoding as well as advances in information technologies and computation have spurred the acceleration of vaccine development. This can be observed in the wide range of technologies and techniques used to develop modern vaccines, which can now target over 25 infectious and non-communicable diseases. Genome-based approaches have enabled the development of vaccines for Meningococcus B or the development of the first ever therapeutic vaccine (for prostate cancer). Similarly, (conjugate) vaccines with multiple antigens or strains now allow for broadened protection while reducing the required number of injections [2][3]. With more than 2.5 million child deaths/year prevented, billions in healthcare cost savings and multiple outcome-related productivity gains, vaccination has become a cornerstone of modern human being of all ages (and financial) health [4]. Looking at vaccine research & development pipelines in Big Pharmaceutical companies, ongoing efforts aim at developing vaccines for more than 50 bacterial, viral, parasitic, degenerative and addictive diseases. This effort includes vaccines against the top 3 killers in developing countries, i.e., lower respiratory infection, HIV/AIDS, and Diarrheal diseases [5].

While population and health professionals rejoice, researchers are faced with mounting challenges hindering the vaccine development life cycle. Among the challenges are: (1) the knowledge disconnect between the disciplines involved – biology, chemistry, engineering, manufacturing, legal, regulator and healthcare – and between stakeholder's views (see Figure 1), which makes the development process very convoluted; (2) the need for sustained capital investment over a lengthy period – hundreds of millions of dollars and 8 to 10 years from research to market with high failure rates – for vaccine development; (3) the costly and stringent storage and handling conditions to be satisfied in order to reduce the loss in vaccine potency and expand an otherwise very short shelf life span (i.e., a year or less); (4) genetic mutations and constantly evolving environment factors making it difficult for certain vaccines to be produced (e.g., HIV-1) [6][7]. A unified, formal, vaccine knowledge-driven approach for the VDLC is needed to enable stakeholders along the VDLC to answer both domain specific and cross-domain questions, quickly, accurately and cheaply, in the context of highly stochastic and complex biological dynamics.

In this project we take a significant step towards a novel knowledge-intensive framework for efficient vaccine development. Our objective is to develop the foundational semantic infrastructure for knowledge and behavior specification, mod-

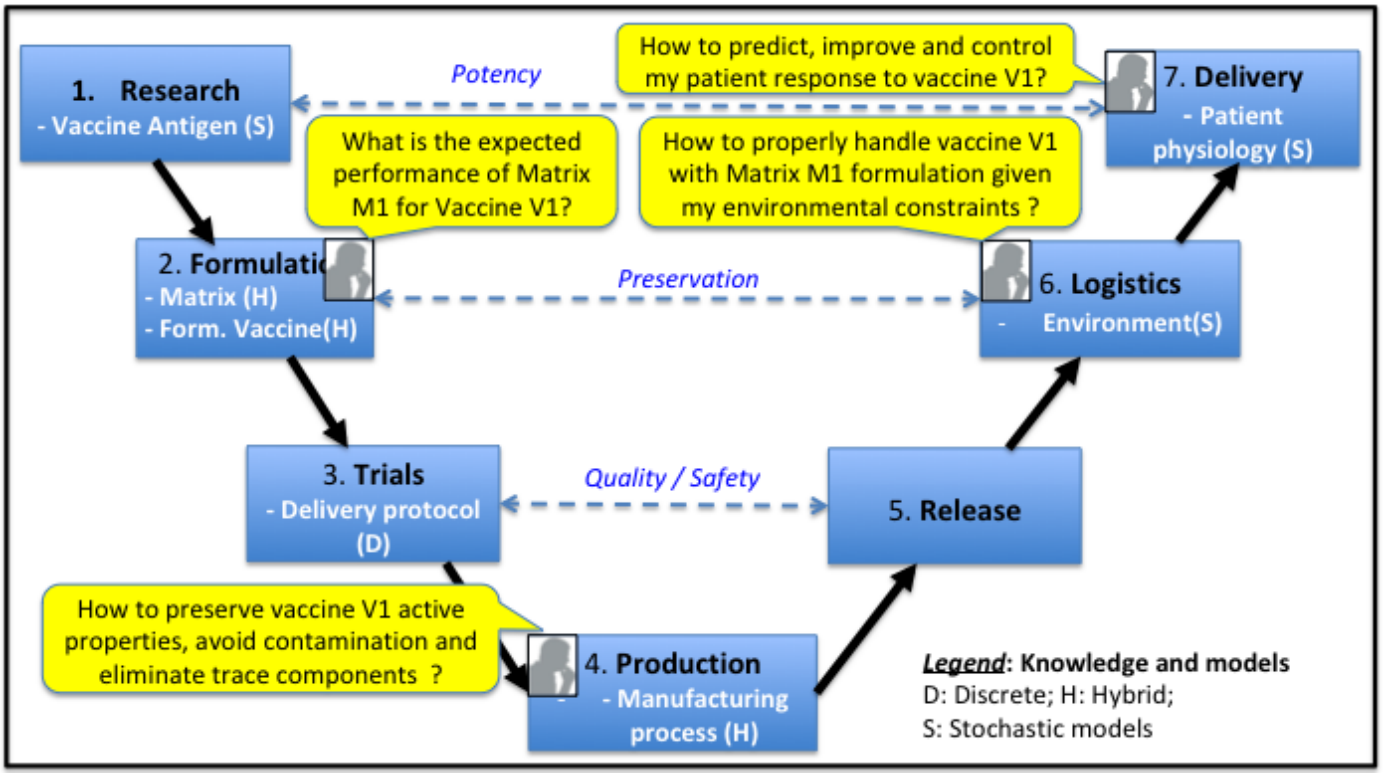


Figure 1. Vee-model of the Vaccine Development Lifecycle and Current Knowledge Gaps

eling and processing across the VDLC as a whole. Therefore, we build on the discipline of systems engineering ability to bridge the gaps between (and integrate) other disciplines, to architect a knowledge-enabled vaccine development platform. The guiding principles of our approach are as follows: (1) formal methods must drive and support the development of vaccine domain models, (2) the latter must properly capture the depth and breadth of stochastic behaviors of biological systems and, (3) models must be reusable, customizable and integrated at will for the purpose of the analysis at hand. The resulting platform supports the integration of biological system dynamic models and, discipline and stakeholder knowledge models thus, enables the emergence of novel architectures, which instantiation can be performed and executed against the requirements of a given application or analysis. Section II is a review of vaccines typology, mechanisms and existing development approaches. Section III introduces mathematical foundations for the formal representation and description of vaccine knowledge and systems biology. Section IV describes the architecture of the framework along with a simplified software implementation infrastructure. An experimental vaccine case study is introduced in Section V to illustrate some of the core capabilities of the framework. The paper concludes with discussions, conclusions and future work.

II. State of the Art: Typology, Mechanisms and Development Lifecycle of Vaccines

Vaccine has been playing a hugely important role in preventing infectious and non-communicating diseases and improving overall quality of living. However, for a vaccine to be successful, (1) its active ingredients should induce an effective and sustained immune response, (2) it must have minimal side effects and, (3) it must be produced cost-effectively at a large scale. Because of the complex nature of vaccine manufacturing it is important to understand and control and or, predict the factors that impacts the efficacy, stability and safety of the vaccine along its process-engineering pathway.

A. Typology and Composition of Vaccines

They are mainly three classes of vaccines. Most conventional or the first generation of vaccines consists of a live, but attenuated form of the pathogen or an inactivated pathogen. *Live, attenuated vaccines* – consists of live viruses that have been extensively passaged through animal hosts until an acceptable balance has been retained between the loss of virulence and retention of immunogenicity. *Inactivated vaccines* – contains microorganisms that have been treated to destroy their infectivity (inactivation). The second generation of vaccines consist only a part of the pathogen – subunit vaccines. *Subunit vaccines* – consists of epitopes around external surface of the pathogen. With recent advances in vaccine science, a third generation vaccines have emerged as DNA and recombinant vector vaccines. *DNA vaccines* consist of non-replicating plasmids, which contain DNA that encodes specific proteins (antigens) from a pathogen. *Recombinant viral vectors vaccine* works by enabling an intracellular antigen expression in the body. Figure 2 illustrates the most common components found in modern vaccines at delivery point. The main components play various functions needed to enable the trigger, execute and maintain host immunization including, (1) elicit and enhance immune response (active ingredients and adjuvants respectively), (2) ensure the stability of various

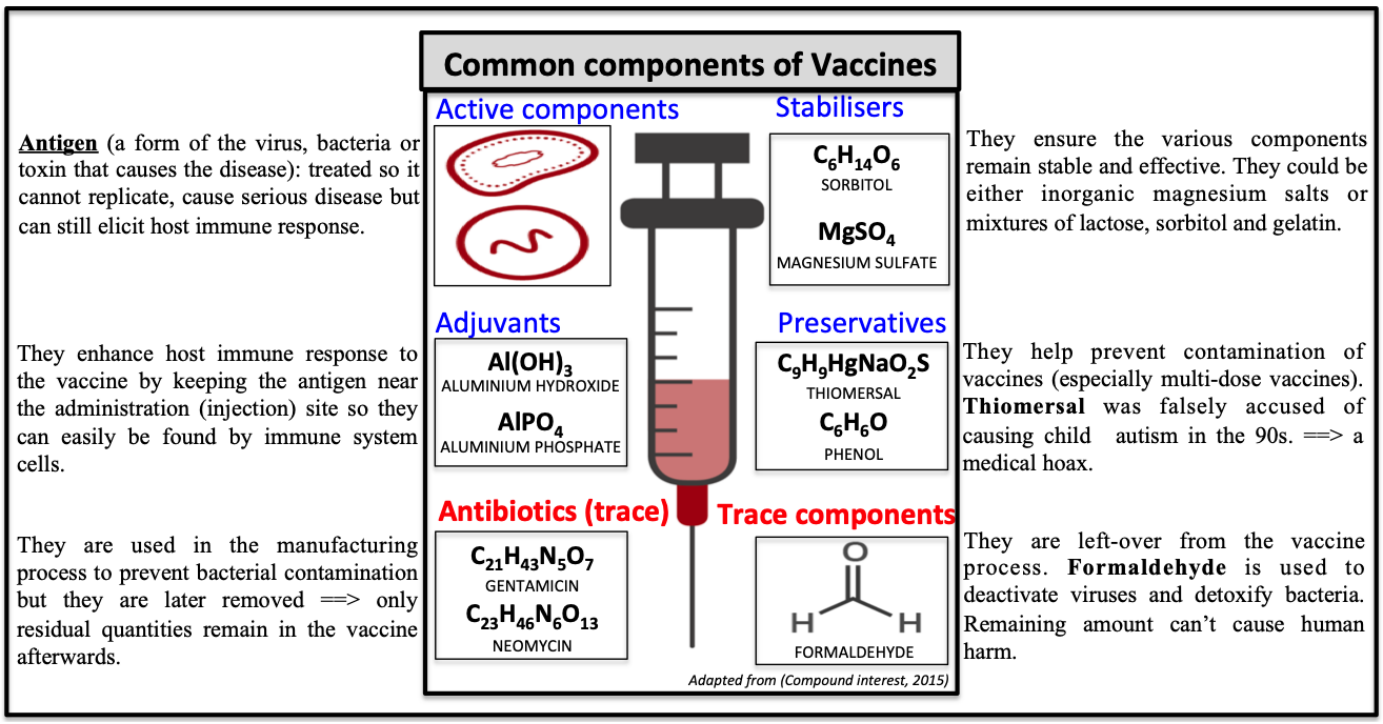


Figure 2. Common components of vaccines(Adapted from [8])

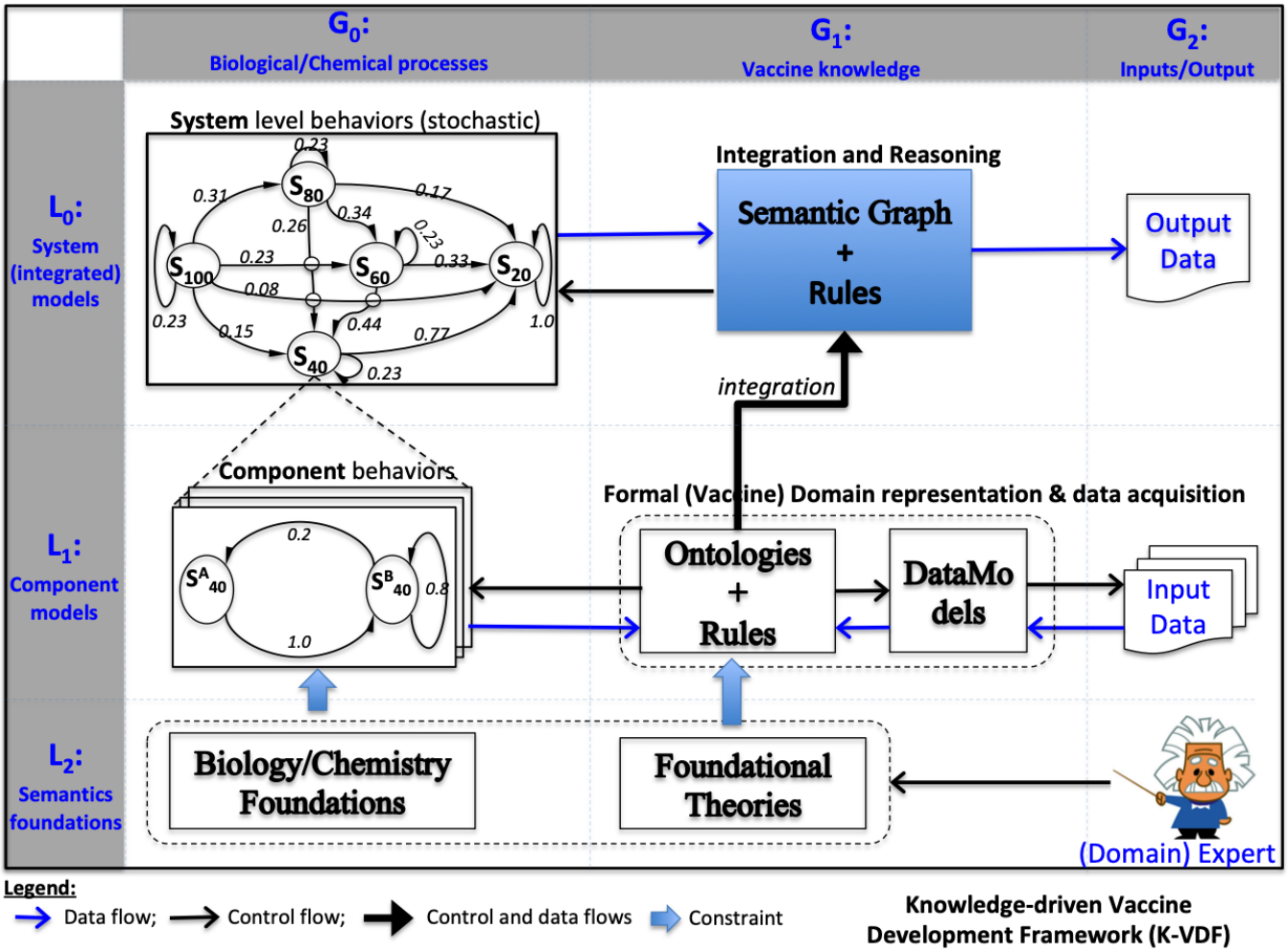


Figure 3. Knowledge-driven vaccine development framework system architecture.

components (stabilizers) and, (3) protect the vaccine against contamination(preservatives). Other components Other inactive elements (e.g., antibiotics and trace components) are inherited from the development and manufacturing process.

B. Mechanisms of Immunization

A vaccine raises immunity through a complex process – yet to be fully understood – in the body. Vaccine protects by inducing immune mechanisms capable of rapidly controlling replicating pathogens or inactivating their toxic components. One immune mechanism is raising antibodies against the vaccine antigen. For example the antigen in an inactivated viral vaccine is the inactivated virus. Vaccine antigens can stimulate a number of cells in the immune system, which includes macrophages, T cells, and B cells. An immune response begins when macrophages ingest the antigen. Fragments of the digested antigen are displayed on the surface of the macrophage. These displayed antigen fragments are recognized by T lymphocytes helper cells, which stimulate B lymphocyte cells to secrete antigen specific antibodies. T helper cells activate killer T cells to actively bind and destroy the antigen. Long-term protection is given by immune memory cells (B and T memory cells), capable of rapidly and effectively re-activating the production of antibodies and killer T cells. Despite the success of three generations of successful vaccines that have eradicated small pox and nearly eradicated polio, there is still a great need for new vaccines and these are emerging far more slowly than we would wish. Successful immunization is not only influenced by various immunological factors (including host physiology and the type and nature of antigen), but also by formulation and delivery aspects.

C. Vaccines Development

The development of vaccines is a complex, lengthy and extremely expensive process involving public, private and non-profit players. It is a high-risk undertaking as many of vaccine candidates fail in preclinical studies. Also, regulatory, technical and manufacturing hurdles lie in the path that translates a vaccine candidate to the final vaccine available in clinics for administration. There are multiples closely coupled stages in vaccines development. The first stage involves candidate selection (exploratory stage) from a fundamental research laboratory – conducting the discovery of antigens – and testing the candidate among animal models. During the pre-clinical stage, development of small case scale material and formulation – a prototype vaccine "medicine" – is done to make material for phase I, II and III studies (clinical development). The exploratory and preclinical can last 1 to 10 years and cost between $10 and $20 millions. Phase I includes test of safety among a sample of 10-100 human subjects to evaluate clinical responses. Phase II focuses on the evaluation of immune responses in a sample of 100-3000 subjects and large scale studies are conducted in phase III to test vaccine efficacy and tolerance. In a clinical development of material candidate, the candidate is cultured, harvested, inactivated (in certain cases), formulated and filled (free dried or in liquid form) in vials syringes packaged and released for distribution. Clinical trials coupled with regulatory approvals can last between 4 to 7 years and are highly capital extensive, with costs in hundreds of millions of dollars.

III. Mathematical Foundations for Vaccine Knowledge Formalization and Systems Biology

A. Knowledge Representation Formalisms

Knowledge representation formalisms are needed to properly capture and formally represent a domain (e.g., vaccine) knowledge as well as reasoning on it. Over the years, researchers have developed several such formalisms including Semantic Networks [9], Frame Systems [10], Description Graphs [11] and Logic-based formalisms [12]. The declarative part of frame systems – a class of logic-based formalisms – are credited for the rise and development of modern Artificial Intelligence (AI) formalisms. Modal and description logics (DL) are descendants of such systems. DL appears to be the most appealing logic-based formalisms for framework like ours, thanks in part to its flexibility of extension to enable complex domain descriptions and the capability to support multi-values attributes and reasoning (for some subsets). Such features are critical to enable the formal representation of a heterogeneous and intricate domain as the vaccine, at various levels of abstractions. Also, we note that some results for description logics were found by translating results from variations of modal logics (propositional dynamic logics, μ-calculus) into description logics [13].

B. Description Logic Semantics and The Semantic Web

Capturing vaccine knowledge and crossing the divide between disciplines along and across the development life cycle depicted in Figure 1 requires mechanisms not just to represent, but also to integrate, share and reuse knowledge across the various stages of the process. Knowledge must (1) be captured, represented in a clear, unambiguous way with respect to the associated domain and the context of use and, (2) lend itself to automated processing and reasoning by machines. This requires data to be enriched and backed by sound semantics to ensure accuracy of facts and inferencing.

Description Logics (DL) formalisms, as fragment of first order logics, provide the sound mathematical foundations and decidability needed to tackle the first part of this challenge [15]. A brief definition of key DL concepts and its $\mathcal{ALC}$ extension are introduced in the appendix of [14]. The strong mathematical foundations of DLs enable the development of machine and human readable ontological languages, such as the web ontology language (OWL), in a systematic way. OWL is the language of choice for creation of ontologies, which are engineering artifacts specifying the intended meaning of a vocabulary used to describe a given domain (e.g., vaccine). As such, ontologies provide explicit semantic meanings that enrich the way models can be branched and integrated across domains of knowledge automatically. Understanding the intricate relationships spanning the vaccine domain and their ultimate affects on vaccine effectiveness will greatly benefit from these capabilities. In [14], $\mathcal{SHOIN}$ and $\mathcal{SROIQ}$ DLs (respectively mapped to OWL1-DL and OWL2 DL) have been identified as appropriate logic-based formalisms for knowledge-driven frameworks such as the one introduced in this work. The computational decidability of OWL2 DL makes it a suitable language for the development of ontologies in our framework.

The second part of above-mentioned challenge can be addressed using semantic web technologies integrated with

reasoner through Application Programming Interfaces or API (e.g., Jena). Semantic web technology resources are organized as a stack, where technologies such as the eXtensible Markup Language (XML), the Resource Description Framework (RDF) and OWL provide the necessary foundations needed by the one on the top, in hierarchical layers. The stack enables the implementation of reasoning that can prove whether or not assertions in the knowledge base are true or false in almost real-time (decidability). Therefore, semantic web technologies are the mean by excellence for automated processing and reasoning over a high variety of distributed and heterogeneous across-domains information such as the ones encountered in vaccine development. Specifically, the various sources of information will be organized, formalized and merged accordingly using semantic models (ontologies, rules and computation extensions) and reasoning will be performed to answer simple and complex biological and/or engineering questions.

C. Stochastic Modeling of System Biology

Beyond the formal description of their structure and properties, the effective capture of the behavior of biological systems (e.g., vaccine antigen, host physiology) across the VDLC is needed to accurately represent and understand the essence of unfolding biological (and underlining) chemical processes at various level of abstractions. Therefore, there is a need for models that can allow for the simulation of the system behavior over time, propagate and predict changes from interactions within systems and with the environment. Researchers have developed and introduced various modeling schemes of biological phenomenons and systems with emphasis on aspects such as body metabolism, neuronal systems, genetic networks or processes (e.g., intracellular processes). Resulting models work fine for cellular level analyses and studies but they are ineffective higher levels of abstractions (e.g., tissue, organs) biological phenomenons [16]. Thus, in order to address those limitations, we opt for a more general formalism – Markov models – in our framework. Such models have been shown effective, in previous work, in modeling and predicting the behavior of highly stochastic biological [17] and biomedical systems [18]. Moreover, they are domain independent and well-suited for integration through segmentation mechanism to domain specific models. In this work, we will use *Markov chain (MC)* formalism to represent actual biological or chemical behavior as a network of states as nodes and directed edges representing allowable transitions between states annotated with their probability of propagation. The graph on the top left corner of Figure 3 – illustrates such MC model. In MCs, feedback and steady-states are allowed as long as all propagation probabilities at each state sum up to 1. A variant of Markov models – *Hidden Markov models* – extend MCs and are suitable for observed system performance (e.g., lab experiments) studies. Markov models, when properly developed and analyzed, are powerful for analysis and prediction of complex system behaviors.

IV. System Architecture and Software Infrastructure

In this section, we introduces and briefly describes the architecture of the proposed framework at the core of effort towards for efficient vaccine development. It is built on top of the mathematical foundations introduced in Section III applied to the vaccine domain knowledge as introduced in Section II. Also, it mirrors a simplified software infrastructure that can enable its deployment at increasingly higher scales and levels of complexity.

A. Overview

The system architecture consists of modules to be assembled as per the needs of the analyses as illustrated in Figure 3. The modules lie at the intersection of three groups of vaccine knowledge categories and three layers of abstractions mirroring various levels of representation of the system. In the first group (G_0), knowledge of component and system biological/chemical dynamics constrained by relevant corresponding (abstract) foundational theories is captured and represented using Markov chains. Knowledge in the second group (G_1) is mostly the formal representation of vaccine, other related domains (e.g., gene, DNA) and foundational fields (e.g., time, space) knowledge as constrained by the corresponding theories. The last knowledge group (G_2) comprises the actual problem input data, the semantically enriched output data resulting from the analysis as well as structured and unstructured (domain) expert knowledge. In the knowledge-intensive framework, not all knowledge types or groups are created equal. They interact with each others – each playing different role – within and across groups to enable the desired functionality of the framework through its analysis-oriented configuration. The main layers of the infrastructure where the various modules are assembled are as follows.

B. Semantic Foundation Layer (L_2)

It provides the mathematical foundations needed by models to ensure effective and unambiguous description of both the domains involved in the analysis and biological/chemical phenomenons. We distinguish foundational theories for known cross-cutting domains (such as time, physical quantities or communication) in module (L_2, G_1) from laws governing biological and chemistry processes in module (L_2, G_0). The Allen Temporal Interval Calculus (ATIC) is a well-suited cross-domain theory that has been shown effective for formal description and reasoning in the temporal domain [14]. In the absence of a valid theory to support the formal description of a domain in the framework, well-accepted domain standards (e.g., CDC Standard for Adult Immunization Practice) as well as heuristics and expert knowledge can be used to fill the void. This offers the possibility for the modelers to inject new theories in the framework for test or evaluation purposes and assess their effectiveness or suitability for given family of problems/analysis. However, the scope and depth of knowledge to be used depends on the application of interest and the goals pursued by the modelers/researchers.

C. Component Layer (L_1)

The component layer enables the modeler to makes use of the formalisms provided by the semantic layer below (i.e., L_2) to create and manage domain knowledge and behavior models that can be reusable across applications. In the context of the VDLC, the knowledge (see module (L_1, G_1)) can be organized and classified in three categories based on their function in a modular way. Core domain (e.g., vaccine antigen, host) knowledge is segregated from cross-cutting domains

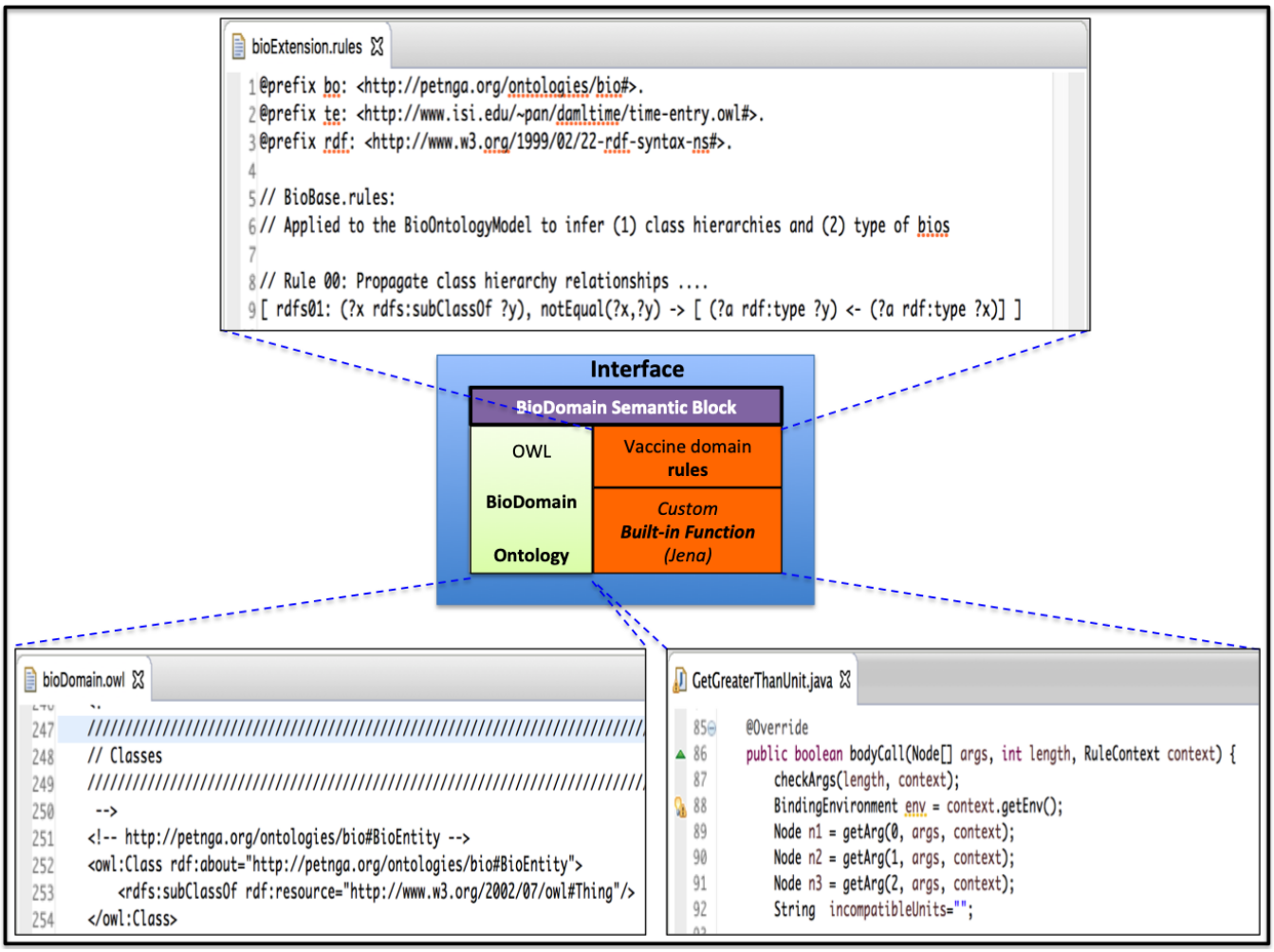


Figure 4. Schematic of a semantic block for the formal description and reasoning about the Vaccine (bioDomain).

(e.g., storage condition, vaccine schedule) knowledge, which in turn, is separated from foundation domain (e.g., time, physical quantity) knowledge. DLs provide the formalisms needed by core domains knowledge while theories such as the ATIC will constraint models of some cross-cutting domains (e.g., vaccine schedule, clinical trial planning). Each domain knowledge in encoded into a "Semantic block" that encapsulates (domain) knowledge in a formal and well-defined manner. Each of the blocks is made of, (1) a domain ontology, (2) set of domain rules, (3) custom computation functions and, (4) interfaces that enable communication between semantic blocks as illustrated in Figure 4. The built-in functions are the glue linking the ontologies to specialized computation platforms and Markov models of system biology (in module (L_1, G_0)) via domain rules as encoded by the reasoner's rules engine. Data-models are templates interfacing input data and ontologies. They enable the modeler to draw from the problem's data stored in input files (module (L_1, G_2)) then, populate the ontology with initial facts in an accurate, systematic and traceable manner. This modular approach adds further rigor and flexibility in the ability of the modeler to build complex applications using reusable semantic blocks (as composite knowledge model).

D. System Layer (L_0)

Leveraging the capabilities of the framework requires bringing together its various modules and pieces in an organized but systematic way. This is needed to close the knowledge gaps between disciplines and stakeholders in along the VDLC as discussed in Section I and answer increasingly complex questions as pictured in Figure 1. Therefore, two tasks need to be performed, i.e., (1) integrate various domain specific knowledge at level L_1 on both the semantic and stochastic behavior sides and, (2) link them and configure the framework accordingly to emulate system level behavior for the application under consideration. Next, the resulting semantic graph (module (L_0, G_1)) is transformed as rules – integrated to stochastic models of the system behavior (module (L_0, G_0)) – are fired. Here, a linkage between the system and component level behaviors to ensure consistency in representations. An "integrator" semantic block can be used as a "semantic controller" that encodes defined system metrics whose instances are checked against system requirements (as constraints). Given the complexity of the integration task, advanced computation capabilities – for controlled and systematic assembly of the models as well as simulation and output generation – such as the ones provided by the Whistle

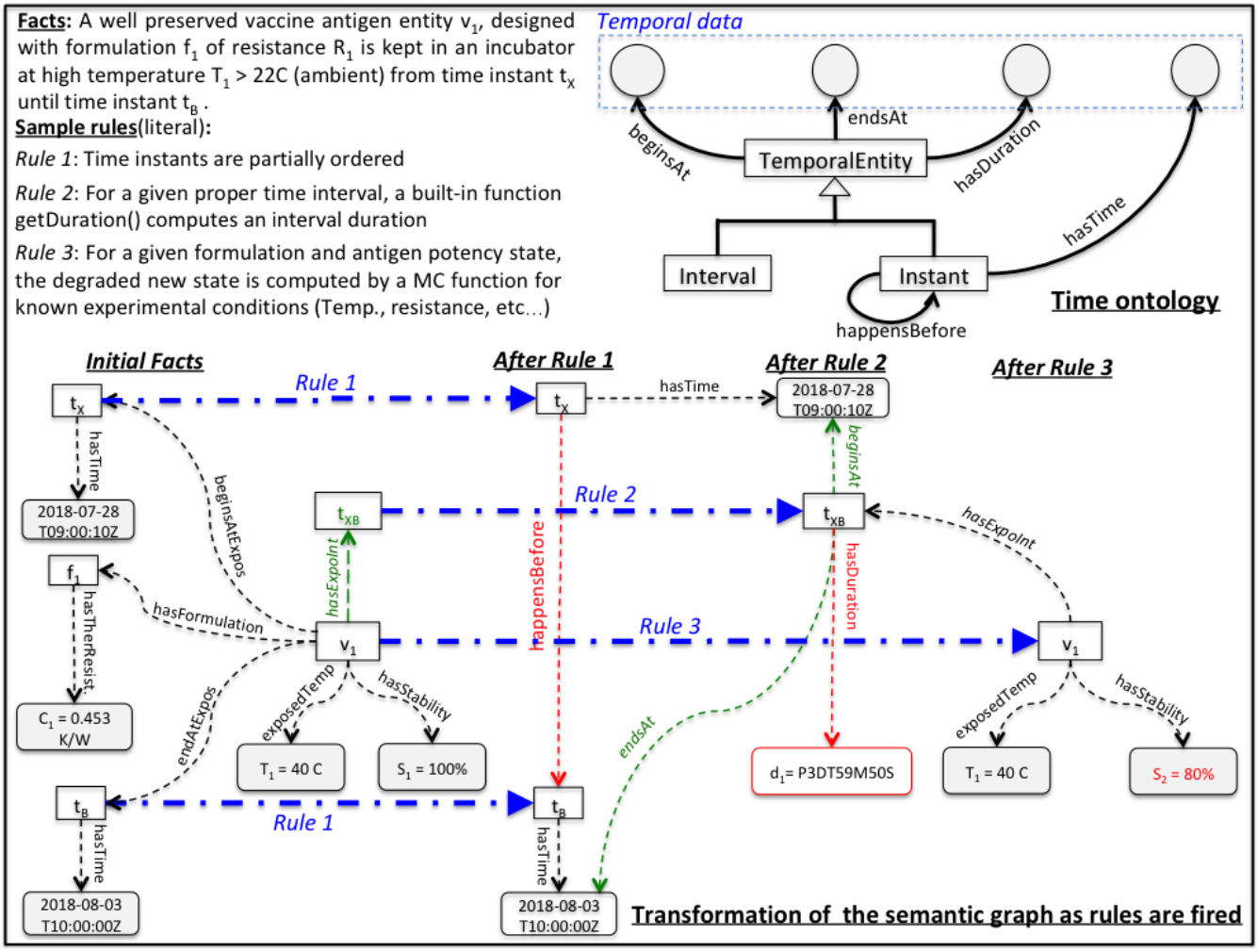


Figure 5. Illustration of rule-based reasoning for vaccine preservation using the temporal domain.

scripting environment are needed. Whistle [19] is a tiny scripting language where physical units are deeply embedded within the basic data types, matrices, branching and looping constructs, and method interfaces to external object-oriented software packages. It is designed for rapid, high-level solutions to software problems, ease of use, and flexibility in gluing application components together. Computational support is added, enabling the language to handle input and output of model data from/to files in various formats (XML, Java, etc.).

E. Working Example of the Usage of the Framework

We describe in this section a configuration and usage of the framework in a scenario where a researcher investigating formulations at step 2 of the VDLC (see Figure 1) looks ahead in step 6 for an answer to the question: *"What is the expected preservation performance of a Matrix M_1 for a vaccine v_1 currently under study ?"*. To that aim, (s)he must leverage the infrastructure of the framework as pictured in Figure 3, in a step-by step configuration and assembly of the various modules as required by the needs of the study then, perform an analysis of the results. In this case, the query is subject to three simplifying assumptions: (A_1) the matrix M_1 (i.e., mainly stabilizers in the formulation), has been properly characterized and the "degradation resistance" C_{f_1} of the formulation is known, (A_2) there are no significant or unknown biological/chemical phenomenons not captured in the framework, and (A_3) computations and reaction times are negligible. The step-by-step details are as follows.

(i) The researcher prepares the input data (e.g., XML file) of the problem as per the framework predefined DataModel: formulation's unique id ($\mathtt{f}_1$), degradation resistance ($\mathtt{C}_{\mathtt{f}_1}$), expected en-route preservation temperature ($\mathtt{T}_1$), initial known stability level of the formulation ($\mathtt{S}_1$), value of time instants when the exposure to $\mathtt{T}_1$ starts ($\mathtt{t}_\mathtt{X}$), and ends ($\mathtt{t}_\mathtt{B}$).

(ii) The data is loaded into the system and the various ontologies (e.g., Time, BioDomain, etc) are populated with instances, i.e., initial facts, as shown on the left side of Figure 5. These are now statements in individual, separated domain knowledge-based as defined in Section IV-C and in module (L_1, G_1) of Figure 3. Specifically, temporal data (i.e., $\mathtt{t}_\mathtt{X}$ and $\mathtt{t}_\mathtt{B}$) are collected as per the template defined in the DataModel, then are deposited in the time ontology while vaccine and formulation data (e.g., $\mathtt{C}_{\mathtt{f}_1}$, $\mathtt{S}_1$, etc.) populate the BioDomain ontology.

(iii) Rule sets for individual domains and parametrized MC models for bio subdomains (e.g., formulation, vaccine antigen)

are provided by the component layer to be used for lower level integration and computation. When and if needed (not in this case), the MC – encoded as built-in function called by the rules engine when the state of the component is needed for inferencing as illustrated in Figure 4 – computes the state of the component and the result is stored in the knowledge base. As indicated in Section IV-B, ontologies and MCs implement foundational theories and system biology/chemistry laws.
(iv) The system integration (layer L_0) is performed by assembling the semantic blocks (ontology + rules + built-ins). The ontologies are integrated and new entity ($\mathtt{t_{XB}}$) and relationship `hasExpoInt` are created to bridge the bio and temporal domains. However, $\mathtt{t_{XB}}$ at this point is a placeholder for a proper time interval in the terminology box (TBox) of the time ontology.
(v) Rule 1 is fired, resulting in the creation of `happensBefore` relationship between $\mathtt{t_X}$ and $\mathtt{t_B}$. This rule belongs solely to the temporal domain and could have been called and executed in step (iii) too.
(vi) Temporal properties `beginsAt` and `endsAt` of $\mathtt{t_{XB}}$ created in (iv) are populated after rule 2 is fired. This is made possible thanks to the fact that the corresponding values are inputs to the built-in function `getDuration()` that computes the duration of temporal intervals (duration of the exposition in this case). The relationship `hasDuration` is created in the temporal domain to store the result of the duration calculated by the built-in function.
(vii) Rule 3 is fired, resulting in the update of the value of property `hasStability` characterizing the stability of the vaccine antigen v_1. For this result to occur, the system level rules engine must pass (via registered built-in) the parameters (i.e.,$\mathtt{C_{f_1}}$, $\mathtt{d_1}$, $\mathtt{S_1}$, $\mathtt{T_1}$) needed by the MC to compute the new state of the system. As in step (iii), the MC model and the ontology are integrated via the built-in function embedded in the rules engine. The new value (S_2) of the property `hasStability` is the answer to the initial question. This, as well as intermediary results, are stored in an output file (e.g., txt format) to be analyzed further by the researcher.

V. Experimental Vaccine Preservation Study

A. Previous Work and Goal of the Study

In [1], we have illustrated the basic implementation and use of our framework in a simplified Oral Polio Vaccine (OPV) formulation under the set of assumptions listed in Section IV-E. An empirical MC model of the degradation of the vaccine stability S_p ($p \in P = \{20, 40, 60, 80, 100\}\%$) – when exposed continuously to temperature T_j for d_k days – was developed with several parameters. The "degradation factor " k_{ijk}^{tf}, which characterizes the ability of the system to maintain itself in a state S_p under the given experimental set up, was found to be given by Equation (1).

$$k_{ijk}^{tf} = \left[\frac{T_{Max} - T_j}{T_{Max} T_j (100 - C_{f_i})}\right]^{\frac{d_k}{d_{Max}}} \quad (1)$$

where T_{max} is the maximum allowable exposure temperature for the experiment and $C_{f_i} \in (0, 100)$ is the "degradation resistance" of a given formulation. Also, the transitions between states (S_p) were computed as per Equation (2).

$$a_{ijk}^{tf}|_{p,q} = (1 - \Delta_{ijk}^{tf}|_{p,q}) k_a k_{ijk}^{tf} \quad (2)$$

where $\Delta_{ijk}^{tf}|_{p,q}$ is the gap of virulence between a state of stability p and one of stability $q < p$ in P and, $k_a > 0$ is a balancing coefficient allowing the probabilities to sum to 1 as per MC modeling rules.

As pointed out in Section II-B, vaccine and vaccination are complex systems and processes not fully understood yet. The current state of vaccine research and development practices does no provide means to characterize key MC model parameters (e.g., C_{f_1}) or ensure that all relevant phenomenons and interactions are captured in models of system biology at the chosen level of abstraction of the representation. Thus, assumptions (A_1) and (A_2) can hardly sustain real-world applications of the framework. Much needed detailed study (outside the scope of this work) is required to address those challenges. Until that becomes a reality, we will develop and use deterministic models to support the computation and predict of the degradation of the vaccine with the lowest possible margins of error. Such models of system biology must be amenable to a smooth integration with the semantic ones in the framework for usage in real-world applications across the VDLC. Therefore, we will conduct laboratory experiments to satisfy those needs.

B. Overview of the Study and Hypotheses

Vaccine antigens are mostly protein. Thus, we use a protein (enzyme) Horse Radish Peroxidase (HRP) as a vaccine model for the preservation study. To perform the preservation studies we use a commercially available HRP. This 40 kDa protein is similar in size to the popular vaccine mimic ovalbumin. The unique structural features of HRP make it a good model protein for analyzing the influence of various excipient properties on protein stability. Because any conformational or structural perturbations of HRP during storage loss of protein activity this is an excellent candidate to study protein stability. The protein also contains four disulfide bonds and numerous metal-binding sites that attract two divalent calcium ions to bind to the protein as enzymatic cofactors. HRP protein is a metalloenzyme that has a noncovalently bound to a heme prosthetic group at the active site. This allows the protein to catalyze the reduction of hydrogen peroxide to water.

C. Laboratory Experiments Setup and Data Collection

The stability of HRP would be tested in three different temperatures at 22^0C, 30^0C and 37^0C in three different excipient formulations containing varying percentages (1-10 w/v%) of a well established excipients used in commercial vaccine formulations. Formulations were also constituted with a constant amount of preservative neomycin (0.01 w/v%), and adjuvant alum (0.02 w/v%) and the dispersant used was phosphate buffer saline (PBS, 0.25 mM) at pH 7.4. Formulations are as follows : (1) F_1 - 1% $MgCl_2$, neomycin (0.01%) and alum (0.02%); (2) F_2 - 5% $sucrose$, neomycin (0.01%) and alum (0.02%); (3) F_3 - 2.5% $trehalose$, neomycin (0.01%) and alum (0.02%). The amount of HRP added to each formulation (F_1-F_3) was 1.33 μg. The stability of the HRP protein at different temperatures in different formulations was tracked using an analytical fluorimetric redox based assay – Ample Red. The stability of the protein HRP was monitored at regular intervals by a redox based fluorimetric assay. A control formulation of HRP formulated in buffer PBS without excipients was

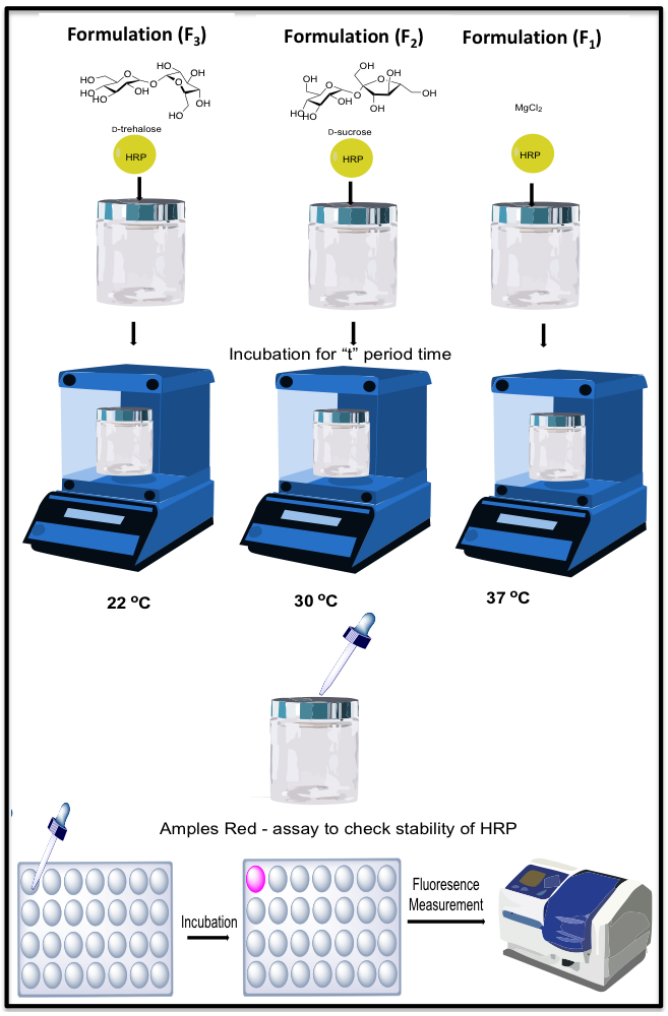


Figure 6. Analysis of the stability of HRP protein in three formulations

kept at -20^0C (FC). Formulations (F_1-F_3) were sealed in 20 mL borosilicate vials kept at the three temperatures mentioned above (22^0C, 30^0C and 37^0C). At regular intervals 50 μL of each formulation with HRP was taken out of each vial added to 50 μL (0.2 mM) of assay reagent Amplex Red in a micro-centrifuge tube. After thoroughly mixing 50 μL was aliquoted and added to 50 μL (4 μM) aliquot of H_2O_2 in a 96 well Costar clear polystyrene plate. The plate was incubated for 1 hour under dark and fluorescence emission was obtained at 590 nm after excitation at 530 nm using a Molecular Devices (M_2) plate reader. The emission intensity of each formulation (F_1-F_3) was recorded and compared to the control (FC) at -20^0C and percent degradation of HRP was calculated for F_1-F_3 at the three different temperatures 22^0C, 30^0C and 37^0C.

D. Experiment Results, Analysis and Limitations

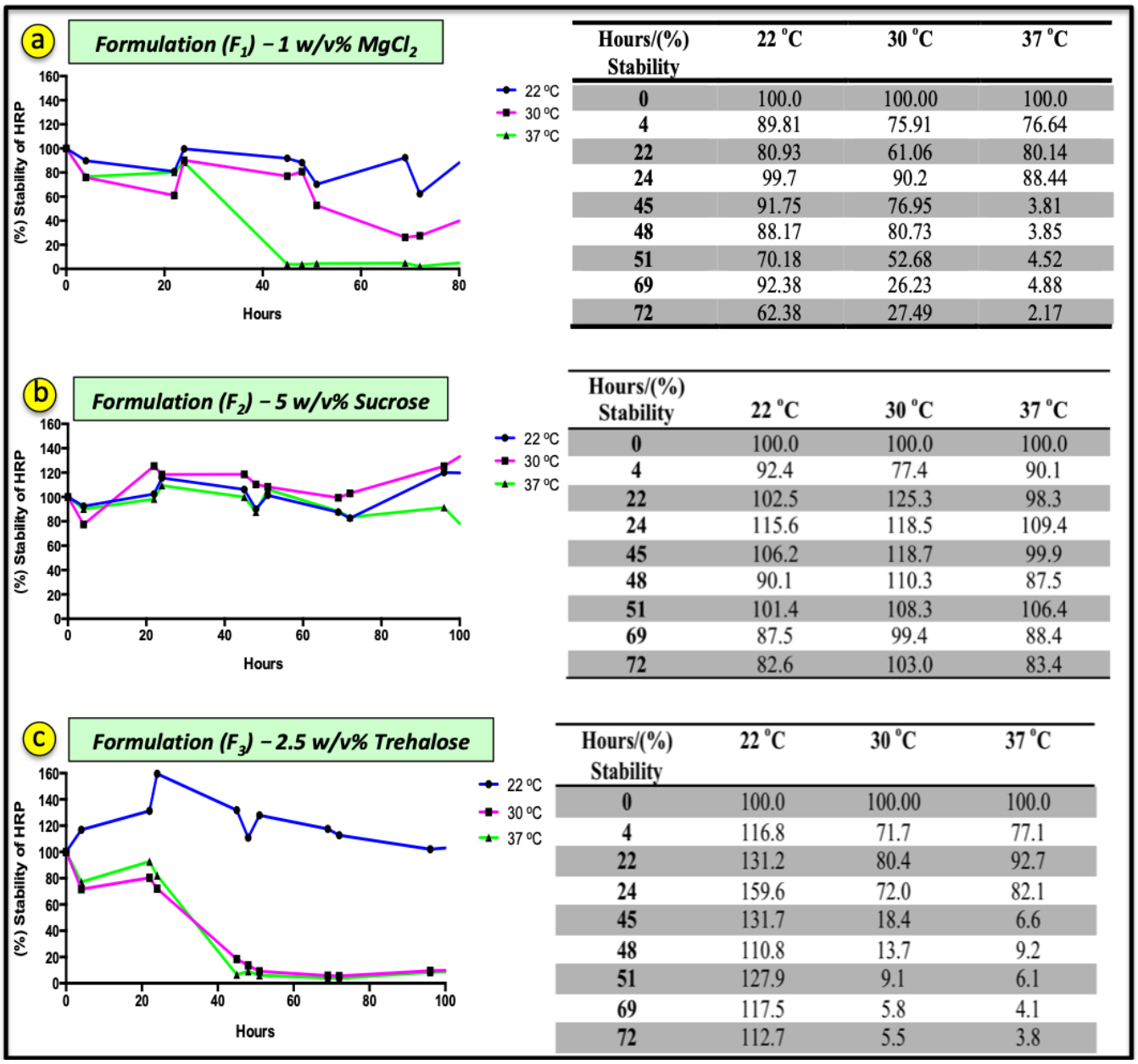
A wide variety of protein stabilizing excipients are used in vaccine development for enhancing the stability of vaccine protein antigens and they are referred to as stabilizing excipients. These excipients have been reported to stabilize the structure of native proteins at moderate (1 w/v%) to high concentrations (30 w/v%). Carbohydrates excipients (i.e., sucrose and trehalose) and polyols (i.e., mannitol, sorbitol) are often used to stabilize protein antigens and protect them from aggregation during lyophilization. Carbohydrates are known to be highly effective in increasing the melting temperature (Tm) of proteins, preventing them from denaturing. Among sugars, sucrose and trehalose have been the most frequently used in thermostabilization. Even though HRP stability data (see Figure 7), at a glance looks erratic, a general trend can be perceived that at higher temperatures (30^0C and 37^0C), $trehalose$ and $MgCl_2$ fail to stabilize the protein HRP. But looking at the stability data of HRP with sucrose (see Figure 7(b)), it is apparent that sucrose could stabilize HRP even at elevated temperatures. Carbohydrates like sucrose and trehalose have high glass transition temperatures (Tg) are known to be more effective in thermostabilizing proteins than other excipients. Salts (i.e., $MgCl_2$) affect in widely different manner when stabilizing proteins. For example, at low concentrations, they could stabilize proteins through non-specific electrostatic interactions while at high concentrations, salting in or salting out would occur. Salting in would preferentially stabilize the protein while salting out would destabilize the protein. At low concentration the hydrated forms of the divalent cation Mg^{2+} has been known to bind to the peptide units through stabilizing hydrogen bonds. Looking at the stability data of HRP protein with $MgCl_2$ compared to the carbohydrate $sucrose$, $MgCl_2$ has provided very little stability.

E. Regression Analysis: Procedure and Results

In the absence of means to properly identify parameters characterizing individual formulations as needed by the MC stochastic model described in Equations (1) and (2), we seek to develop indirect means for predicting the degradation of a given formulation. Thus, we formulate and construct regression models correlating (statistically) independent experiment variables (introduced in Section V-C) and the percentage of degradation of the protein (i.e., our surrogate vaccine antigen) as response. The variables considered for this analysis are primary the temperature at which the formulated protein is exposed to (x_2) and the duration of exposition at that temperature (x_1) for formulations $MgCl_2$ and $sucrose$. In the case of $trehalose$, a third variable – the percentage of stabilizer (x_3) in the formulation solution – is added to the mix. We use the data collected in Section V-C to perform the analysis. Simple and multiple regression models accounting for the variables individually or together and their interactions are constructed and identified using the following nomenclature.

$$M_l^{j.k} \models Y = f(x_1, x_2, x_3) \quad (3)$$

where $M \in \{L, Q, P\}$ is the regression model, i.e., Linear(L), Quadratic(Q) or Polynomial(P) of order 3 for the formulation under study; $j \in \{1, 2, 3\}$ is the type of the formulation, i.e., $MgCl_2(1)$, $sucrose$ (2) and $trehalose(3)$; $k \in \{1, 2, 3, 4, 5, 6\}$ is the percentage of the stabilizer in the formulation of interest, i.e., 1%(1), 2.5%(2), 5%(3), 10%(4), 20%(5), and a combination of several percentages(6). Also, $l \in \{1, 2, 3, 4\}$ is the temperature at which the protein is exposed, i.e., average room temperature of $22^0C(1)$, high temperature of $30^0C(2)$, body temperature of $37^0C(3)$ and All temperatures(4). Even though this representation allows us to cover all configurations of regressions, we will be focusing on ones enabling us to capture, represent and identify multiple regressions in a very unique ways. Thus, $l = 4$ in such models.



Hours/(%) Stability	22 °C	30 °C	37 °C
0	100.0	100.00	100.0
4	89.81	75.91	76.64
22	80.93	61.06	80.14
24	99.7	90.2	88.44
45	91.75	76.95	3.81
48	88.17	80.73	3.85
51	70.18	52.68	4.52
69	92.38	26.23	4.88
72	62.38	27.49	2.17

Hours/(%) Stability	22 °C	30 °C	37 °C
0	100.0	100.0	100.0
4	92.4	77.4	90.1
22	102.5	125.3	98.3
24	115.6	118.5	109.4
45	106.2	118.7	99.9
48	90.1	110.3	87.5
51	101.4	108.3	106.4
69	87.5	99.4	88.4
72	82.6	103.0	83.4

Hours/(%) Stability	22 °C	30 °C	37 °C
0	100.0	100.00	100.0
4	116.8	71.7	77.1
22	131.2	80.4	92.7
24	159.6	72.0	82.1
45	131.7	18.4	6.6
48	110.8	13.7	9.2
51	127.9	9.1	6.1
69	117.5	5.8	4.1
72	112.7	5.5	3.8

Figure 7. Stability of HRP at temperatures 22^0C, 30^0C and 37^0C (a) in Formulation F_1 - 1% $MgCl_2$, (b) in Formulation F_2 5% $sucrose$ and, (c) in Formulation F_3 2.5% $trehalose$ for a period of 72 hours

We follow a rigorous data analysis process to generate and ensure the quality of the resulting regression models for the each of the formulations. First, the data is cleaned from outliers using 95% confidence interval for the response. Second, the data is checked for confounding to make sure the independent variables are independent from one another. This is done by plotting scatter plot for pairs of independent variables and checking for collinearity. In this study, no such relationship were found in any of our formulation data sets. With those foundations in place, the next move is to fit response surface models to individual formulation stability percentage (%HRP), as a function of controllable factors x_i, ($i \in \{1, 2, 3\}$) as defined above. One model of each of the three types, i.e., Linear(L), Quadratic(Q) or Polynomial(P) is created. Plots of residuals versus fitted values are generated and used each time to check for violation of assumptions for error in regression models. This exercise has helped us uncover inconsistencies in the regression models for the $trehalose$ formulation. However, applying a logarithmic transformation to the data has resulted in more normal (randomly distributed points) residual plots, and better regression models. Finally, the best model for each of the formulation was selected by comparing models' coefficient of determinations or goodness-of-fit statistic (i.e., R^2). Table I summarizes our findings.

Polynomial models' coefficients of determination (R^2) are the highest of all model types for all formulations. Thus, they are the best fitted models for the response as shown in Table I. We note here that the response function for the $trehalose$ is

TABLE I. Selected multiple regression models for each candidate formulation. The % of trehalose in model $P_4^{3.6}$ comprises 2.5%(2), 10%(4) and 20%(5)

Formulation	Model	Coef. det.(R^2)	Response function: $Y = f(x_1, x_2, x_3)$
$MgCl_2$	$P_4^{1.1}$	82.91 %	$Y = 96 + 1.90x_1 - 0.8x_2 - 0.0499x_1^2 + 0.021x_2^2 + 0.015x_1x_2 + 0.000249x_1^3 + 0.00066x_1^2x_2 - 0.00244x_1x_2^2$
$sucrose$	$P_4^{2.3}$	72.36 %	$Y = -12.9 + 0.46x_1 + 7.88x_2 - 0.0299x_1^2 - 0.135x_2^2 + 0.118x_1x_2 + 0.000115x_1^3 + 0.000206x_1^2x_2 - 0.00219x_1x_2^2$
$trehalose$	$P_4^{3.6}$	88.94 %	$ln(Y) = 9.39 + 0.4388x_1 - 0.337x_2 - 0.254x_3 - 0.003164x_1^2 + 0.00582x_2^2 + 0.01464x_3^2 - 0.02049x_1x_2 - 0.00820x_1x_3 + 0.0074x_2x_3 + 0.000017x_1^3 + 0.000036x_1^2x_2 + 0.000032x_1^2x_3 + 0.000230x_1x_2^2 + 0.000162x_1x_2x_3 + 0.000018x_1x_3^2 - 0.000000x_2^2x_3 - 0.000453x_2x_3^2$

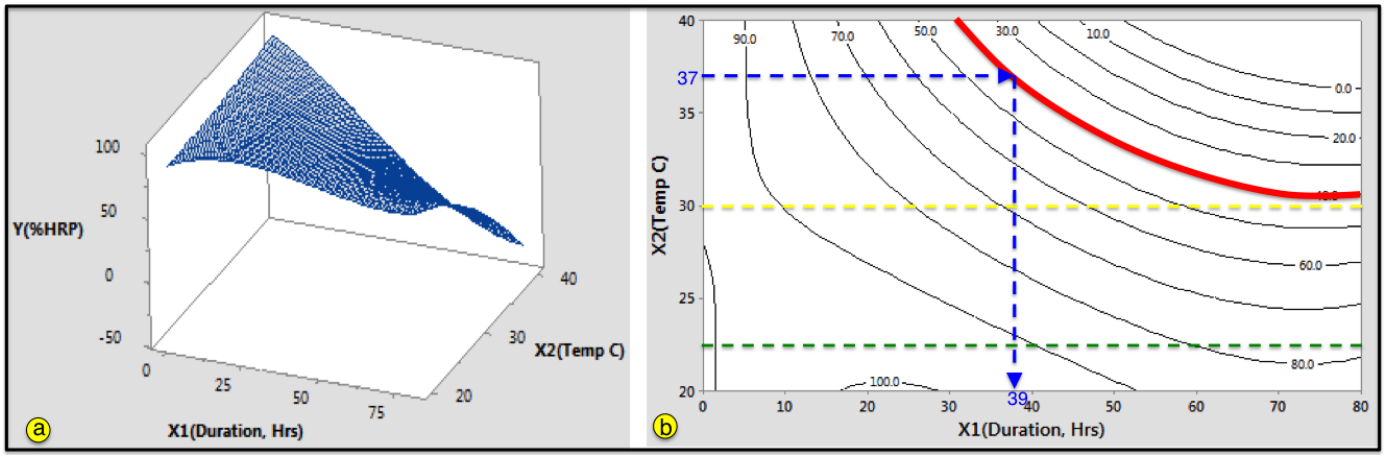


Figure 8. Response surface (a) and contour (b) plots for $MgCl_2$ based on multiple regression model $P_4^{1.1}$

logarithmic thanks to the data transformation and the presence of three – instead of two – independent variables. A plot of the response surface for model $P_4^{1.1}$ ($MgCl_2$ formulation) is shown in Figure 8(a). The hyperplane representing the response, i.e., the percentage of HRP is bended downward as both the duration of exposition and temperature increase. This is consistent with the expected behavior of the system but also as previously found with the empirical model. However, we gain additional insight in the observation that, for this particular formulation, its stability clearly takes a dive as temperature at which the formulated antigen is exposed increases. This suggests that temperature (i.e., intensity of the heat), more than duration of exposure, is the main driver of the breakdown of the stabilization property of the formulated protein. Figure 8(b) is the contour plot for the same response surface. It offers a better view of the devastating effect of temperature on the stability of the formulation. For instance, considering a targeted minimum threshold of 40% stability (as for the Oral Polio Vaccine), one can clearly see that it will take less than 40 hrs (1.6 day) of exposure of the formulated antigen at body temperature (i.e., $x_1 = 37^0C$) to loose 60% of its stability. For the same duration, it looses only 30% of its stability at $x_2 = 30^0C$ and less than 10% stability loss at room temperature ($x_3 = 22^0C$). These results also explain well-established state-of-the-art vaccine preservation practices of keeping vaccines at lower temperatures (i.e., higher logistic costs) to maintain its stability over a long period of time. Finding the formulations that can achieve the same results or better at higher temperatures (i.e., lower logistic costs) remains the holy grail of vaccine preservation research.

VI. Discussions

As of now, it is difficult for real-world applications to sustain assumptions (A_1) and (A_2) stated in the working example of the framework described in Section IV-E. When it comes to assumption (A_1), the case study highlights the challenge of developing accurate and precise system biology models to be integrated with semantic models in our framework as described in Section III and pictured in Figure 3. This is needed to support prediction and reasoning in the framework. In the face of challenges regarding the characterization of formulations to support the full definition of stochastic models (MC), we have developed regression-based models for stability prediction in our prototype implementation on a preservation study. Such models can be used under specific conditions – in lieu of actual MC models – for build-in functions enabling computations such as the ones in rule 3 (see Figure 5). Regression models establish statistical (not causal) correlations between independent variable(s) and a response under specific and well-defined set of conditions. This limits its scope of use and its ability to support the explanation of underlining biological/chemical phenomenon. This will not be resolved until proven and full characterizations of biological agents (e.g., vaccine antigens) are available to be used for models (such as the MC models) used in this framework.

Inconsistencies and out of range results in selected regression models for $sucrose$ ($P_4^{2.2}$) and $trehalose$ ($P_4^{3.6}$) suggests that there are important underlining chemical/biological phenomenons unaccounted for by the model. This is a translation of a clear violation of assumption (A_2). Addressing this problem will require uncovering such phenomenons for the given formulation followed by the identification of explanatory variables to be tracked during experiments and

refinement of current regression models to account for the new variable(s). The complexity of the problem significantly increases if we consider that, to date, there are 380 established antigen stabilizing compounds or generally-regarded-as-safe (GRAS) excipients candidates that could possibly be used in a vaccine formulation [20].

The ability of the framework to survive assumption (A_3) in real-world applications depends on the capability of the underlining software infrastructure supporting implementation. Even though one running VDLC-related applications using the framework would not want them to last for ever, they are not safety-critical. Thus, real-time computations are not a "must" but, fast computations – especially when faced with large volume and heterogeneous data – are needed. As pointed out in Section III-B, OWL – the language we use to develop semantic models in this framework – enables both human and machine processing of vaccine and domains knowledge over the World Wide Web (WWW). Proper integration with databases, web-based interfaces, and cloud computing as well as with the appropriate configurations, fast, integrated yet distributed solutions are possible. Therefore, both batch and streaming-based processing of data through the framework are possible.

VII. Conclusion and Future Work

In this paper, we have introduced and described a knowledge-intensive framework for behavior specification, modeling and reasoning for efficient vaccine systems engineering. This research is motivated by limitations of state-of-the-art vaccine development approaches in capturing, representing and reconciling domains and disciplines knowledge and viewpoints across the vaccine development life cycle in an effective manner. The inherent highly stochastic behavioral nature of biological elements such as vaccines coupled with knowledge disconnect between stakeholders (e.g., chemists, biologists, clinicians, the public, big pharma, etc.) with sometimes conflicting interests add to the numerous technological challenges of engineering such complex biological systems. This leads to long, complex, and costly efforts with high failure rates currently observed in vaccine development initiatives. Also, potency of successful vaccines is difficult to predict and very expensive to preserve in the face of changing and challenging environmental conditions as well as limited resources.

The knowledge-intensive framework is shown to be a possible solution to successful vaccine systems engineering moving forward. Description logic semantics provide the necessary formalisms needed to capture, represent and reason about vaccine and foundational domains knowledge in a clear, unambiguous way with respect to the associated domain and the context of use while enabling automated processing and reasoning by machines through semantic web technologies. Library of reusable semantic components – i.e, semantic blocks comprising an ontology, rules, computation and communication interfaces – encapsulate knowledge in a formal and well-defined manner. They are integrated to stochastic models of vaccine system biology – Markov Chains (MC) – of the underlining unfolding biological and/or chemical processes at various level of abstractions throughout the development lifecycle of the vaccine, when needed. This layered and modular structure enables flexibility in the assembly – via integration – of models of various level of complexity and types in support of the investigation of research issues that cut across domains in the vaccine development lifecycle. Thus, this will help bridge the gap between domains and stakeholders along the development lifecycle, with the ripple effect of shortening the development cost, length and complexity.

A step-by-step implementation procedure coupled with a prototype vaccine preservation study have shined some light in the implementation of the framework. Such studies, if successful, can replicate actual preservation conditions in extreme weather (e.g., subsaharan Africa) and guide the design and selection of the most effective formulation able to stabilize the vaccine in given situations. However, limitations in the current state of vaccine research and development practices in, (1) providing means to characterize key MC model parameters and, (2) ensuring that all relevant phenomenons and interactions are properly have appeared to be a challenge to the proper account of system have stood on the way. Thus, we have designed and conducted laboratory experiments, which coupled with regression analysis of stability data has resulted into multiple regression models that were used as an alternative path. The resulting deterministic models are shown to provide statistically significant and satisfactory results under the specific set of experimental conditions. Responses surfaces and contour for on the "on-the-fly" prediction have been produced.

Future work needs to address challenges related to the full and accurate characterization of vaccines properties standing on the way of the creation of stochastic models of biological elements (e.g., antigens) to be used in the framework. Therefore, advanced laboratory experiments are needed to that aim but also to uncover and understand relevant phenomenons of interest contributing to the system response (e.g., stability of the vaccine). The application of the appropriate design of experiments will be needed to ensure cost effectiveness and overall efficiency in studies and analyses. Bringing the benefits of the framework introduced in this work to day-to-day work of stakeholders across the vaccine development lifecycle will also necessitates further work on the refinement and validation of the framework for various vaccine types, analyses and cross-cutting concerns (e.g., potency, preservation, safety) and various environmental conditions. Finally, the collaborative development of domain and discipline knowledge across the development lifecycle – e.g., vaccine ontology as in [21] – is highly suitable to foster dialogue and synergy between stakeholders.

Acknowledgment

The authors would like to thank the students who contributed to this research with data collection and analysis: Kavini Rathnayake, Unnati Patel, Veera Venkata Naga Manohar Devarasetty, Melinda Mustain, James Johnson and Anoop Kumar Reddy Gudipati.

References

[1] L. Petnga and S. Jayawardena, "Knowledge-driven Vaccine Systems Engineering," The Thirteenth International Conference on Systems (ICONS), Athens, Greece, April 22 - 26, 2018.

[2] H. Stevens and K. Debackere, "Vaccines: Accelerating Innovation and Access," *Global Challenges report World Intellectual Property Organization(WIPO)*, 2016.

[3] A. Loharikar, L. Dumolard, S. Chu, T. Hyde, T. Goodman, and C. Mantel, "Status of New Vaccine Introduction Worldwide, September 2016," 65(41);11361140, US. Center for Disease Control(CDC), October, 21, 2016.

[4] World Health Organization(WHO), "Immunization coverage," WHO Factsheets/Detail, July, 18, 2013.

[5] International Federation of Pharmaceutical Manufacturers & Associations(IFPMA), "Vaccine research and development," IFPMA Resources/Graphics, April, 08, 2013.

[6] The United Nation Children's Fund (UNICEF), "Vaccines: Handled with Care," Division of Communication, New York, USA , 2004.

[7] Pharmaceutical research and Manufacturing of America(PhRMA), "Vaccines Factbook," PhRMA, 2013.

[8] Compound Interest, "A Summary of Common Vaccine Components," Accessible at: https://www.compoundchem.com/2015/02/10/vaccines/; Retrieved: November 17, 2018.

[9] J.F. Sowa and A. Borgida, "Principles of semantic networks: explorations in the representation of knowledge," John F. Sowa (Ed.), 1991.

[10] P.J. Hayes, "The logic of frames," Frame Conceptions and Text Understanding, deGruyter, pp. 4661, Berlin, 1980.

[11] M. Pavlic, A. Mestrovic, and A. Jakupovic, "Graph-based formalisms for knowledge representation," 17th World Multi-Conference on Systemics, Cybernetics and Informatics, July 912, Orlando, Florida, USA, 2013.

[12] F. Baader, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider, "The Description Logic Handbook: Theory, implementation, and applications," Cambridge, 2003.

[13] K. Schild, "Terminological cycles and the propositional with mu-calculus," J. Doyle, E. Sandewall, P. Torasso (Eds.), 4th Int. Conference on the Principle of Knowledge Representation and Reasoning (KR-94), pp. 509520, 1994.

[14] L. Petnga and M. A. Austin, "An Ontological Framework for Knowledge Modeling and Decision Support in Cyber-Physical Systems," Advanced Engineering Informatics, Vol. 30, No. 1, pp. 77-94, January, 2016.

[15] F. Baader, I. Horrocks, and U. Sattler, "Description Logics," In: Handbook of Knowledge Representation, pp. 135 - 180, Elsevier (US), 2008.

[16] B. Ingalls, "Mathematical Modeling in Systems Biology: An Introduction," Applied Mathematics, University of Waterloo, CA, 2012.

[17] C. J. Tomlin and J. D. Axelrod, "Biology by numbers: mathematical modeling in developmental biology," Nature Reviews; Genetics 8: pp. 331 - 340, 2007.

[18] M. Mosteller, M. A. Austin, R. Ghodssi, and S. Yang "Platforms for engineering experimental biomedical systems," 22th Annual INCOSE International Symposium (IS 2012), Rome, Italy, July 9-12, 2012, 2012.

[19] P. Delgoshaei, M.A. Austin, and A. Pertzborn, "A Semantic framework for modeling and simulation of cyber-physical systems," International Journal On Advances in Systems and Measurements, vol. 7, no. 3-4, pp. 223238 (EU), 2014.

[20] U.S. Food and Drug Administration (FDA), "Generally Recognized As Safe (GRAS) food substances," Accessible at: https://www.accessdata.fda.gov/scripts/fdcc/?set=SCOGS; Retrieved: November 17, 2018.

[21] University of Michigan School of Medicine, "Vaccine Ontology," Accessible at : http://www.violinet.org/vaccineontology/introduction.php; Retrieved: November 17, 2018.