Knowledge Integration for Bio-Threat Response

A White Paper

September, 2005

Editors: Karin Verspoor, Cliff Joslyn[1], John Ambrosiano (Los Alamos National Laboratory), and Lynette Hirschman (MITRE)

Contributors: Alex Bäcker (Sandia National Laboratories and Caltech), Olivier Bodenreider (NLM), Peter Karp (SRI), Henry Kelly (Federation of American Scientists), Stephanie Loranger (Federation of American Scientists), Mark Musen (Stanford U.), Ram Sriram (NIST), Chris Wroe (U. Manchester)

Abstract

Bio-threats require rapid analysis and response to prevent widespread consequences to the population. The inability to easily link and exploit biological knowledge for both human and automated analysis is a major limitation on the speed of complex knowledge development and bio-threat response. Today, linkage among the vast array of biological knowledge repositories is primarily by hand. Key requirements for smooth linkage of knowledge sources include shared ***ontologies*** of concepts and semantic relations, capabilities for unifying terminology and extracting meaningful relations from ***text***, and ***inference*** mechanisms to link and unify heterogeneous databases. While research in all these areas continues, the gaps to the needed knowledge integration technologies are immense, and would benefit from co-ordinated cross-agency funding.

[*] **Corresponding Author:** Knowledge Systems and Computational Biology, Computer and Computational Sciences, Mail Stop B265, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, joslyn@lanl.gov, http://www.c3.lanl.gov/~joslyn, (505) 667-9096.

Introduction

Disease outbreak, whether from natural causes or bioterrorism, remains a deadly threat to humankind and to political stability, and evidence is mounting that this threat can be minimized only through timely detection and rapid response. In an increasingly connected world, where emerging diseases can move quickly into new regions (e.g. West Nile Virus) or jump from animals to humans (e.g. SARS, avian influenza), and the threat of terrorist or even military biological attack is of constant concern, the existing limitations on our ability to draw on medical and biological knowledge to analyze emerging diseases constitutes significant vulnerability.

The field of biomedicine has exploded in recent years, with an incredible increase in knowledge and data available to scientific researchers. At present, these are distributed among a plethora of sources and repositories that have only been integrated in a loose, *ad hoc*, and largely manual manner. The inability to easily link and exploit biological knowledge for both human and automated analysis is a major limitation on the speed of complex knowledge development.

It is true that databases are interconnected as never before, and researchers can access a staggering array of databases and document repositories from their desktops. But beyond this point of basic accessibility, users are on their own to know *what* is in each resource and to draw common information together. Whatever efforts have been made to achieve a higher order *semantic* interoperability, integration of the *knowledge* within these databases is halting and surprisingly technology-poor: manual linkage by experts predominates, with automated methods largely relegated to traditional keyword matches and Boolean search.

Multiple efforts are underway to integrate biological *content*, for example the Biodefense Knowledge Center (BKC) at DHS and the Unified Medical Language System (UMLS)[i] at NIH/NLM[ii]. Organism-specific databases, especially for model organisms such as yeast, fruit fly, and mouse, also continue to develop and even integrate, driven by their particular communities. And there are many efforts ongoing for conceptual and terminological unification, such as the Gene Ontology (GO)[iii,iv] and NCBI's MeSH ontology.

Such efforts are necessary and welcome, and there is no doubt that active efforts at content development and unification will continue to be essential. But these databases are becoming large and numerous very quickly. In preparing this paper, we have gathered information documenting the existence of over 275 distinct online databases, covering general protein and nucleotide sequences, and others specific to humans, other mammals, invertebrates, bacteria, and viruses and phages.

Most disturbing is that these efforts continue to be primarily manual and highly labor intensive: where is the *technological* assistance for these researchers? The information technology revolution has, in effect, created this problem by inundating the research community with vast new quantities of data. In turn, *technological* solutions may provide

significant assistance, with the proper, focused investment in the necessary science and technology.

This lack of technological tools for biological researchers poses significant obstacles for our nation's ability to respond quickly and effectively to newly emerging biological threats. There is a serious need for enhanced technology in this area, including tools which facilitate the extraction of semantic information from databases and texts, the establishment of standardized vocabularies and terminologies, the codification of the semantic relations between terms in ontologies of concepts connected by semantic relations, and the merger and linkage of such ontologies once created.

This paper motivates the need for a concerted effort in the area of integration of biological knowledge, both by outlining scenarios with national security implications for which such knowledge integration is a critical component, and by describing the inadequacies of the existing efforts in this vein. We will detail the potential military significance of this work, and suggest a path forward to addressing the knowledge integration problem through a future program in technology development for knowledge integration.

The Need: Bio-Threat Response

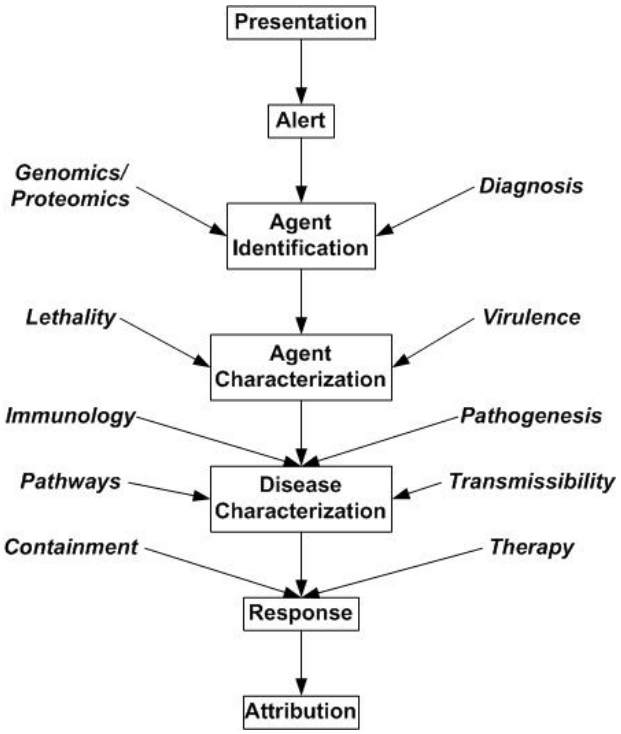
Today's soldier faces myriad potential threats, including from biological weapons. The following scenario outlines how ontological technologies can make it possible to reduce the time of identification and response in the case of a biological weapons attack on our troops. The basic response scenario is also applicable to the case of the outbreak of a new virus or disease within the general population.

A group of soldiers in Iraq suddenly exhibit acute, flu-like symptoms. Worse, the illness is spreading, threatening unit cohesion. Simple diagnostic tools at the base are inconclusive; diagnosis and treatment is unknown. Instantly the field commanders are faced with critical questions: What is the nature of the pathogen (bacterial, viral)? Is there an index case? Is it similar any known organisms (perhaps by symptomatic evidence)? Is it presenting a *mixture* of symptoms? *Who* knows about this kind of disease? Who should be *alerted*? What is the transmission path (phases and vectors)? How do we contain it? How do we treat it? The necessary analysis is guided by a workflow like the one suggested in the figure below.

Biological samples from the infected soldiers are flown to a molecular biology lab along with the patients' medical records (including differential diagnosis, test results and procedures performed). Upon receiving the samples the laboratory technicians begin identifying and characterizing the biological agent infecting the soldiers. The microbiologist would isolate and culture the pathogen. Once isolated, the next step is to characterize the pathogen. The diagnosis is mainly based on DNA/RNA sequence homology search, by direct sequencing followed by database search or by hybridization assays such as DNA microarrays.. The pathogen genetic material may be amplified for sequence homology search through PCR (Polymerase Chain Reaction). Once the DNA is sequenced, database searches can help to determine the type of pathogen, and its potential

homologues. The searched databases might include Genbank or Swissprot and make use of a BLAST search. Once a potential pathogen or homologues are identified, a Pubmed search of the biomedical literature would be performed to understand what is known about the pathogen and its homologues.

While the pathogen is being identified and characterized in the laboratory, epidemiologists work with clinicians to understand how the pathogen is spreading and the characteristics of the disease. This work is aided by knowing what the pathogen is, or to what it is related. Immunologists would investigate the host response to infection with this pathogen and find ways to boost host response to eliminate the pathogen. Animal models, though difficult to develop, can be useful in therapeutic trials; model organisms can potentially be used testing in initial tests of therapeutics. Once the pathogen is identified and its mode of infection is known, a response can be coordinated. Therapies are administered and infected soldiers quarantined, hopefully containing the outbreak.



Of course, we have actually had similar experiences, albeit not in a battlefield context: hanta virus, SARS and West Nile virus are just a few examples of recent outbreaks of novel, or at least not initially recognized, pathogens, and there is much current nervousness about avian flu, Dengue fever, and, of course, potential military and terrorist threats.

In any particular incident, if done properly, the steps identified in this workflow would take months. Lessons learned from SARS include that the lack of early warning allowed rapid spread; that an outbreak can rapidly have global reach and devastating economic impact, in the billions of dollars; airplanes are a new vector; international coordination is possible, for example for collaborative sequencing efforts; and containment through isolation worked, even though we still have no drugs or vaccines.

Information sharing was critical in the fight against SARS. Generally in such outbreaks, multiple databases are consulted by different communities with little, or more likely no, overlap. The clinicians might use a differential diagnosis database such as Gideon[v], a clinical database, and Pubmed. The molecular biologist will consult several genomic databases (Genbank, Swissprot, Expasy, etc.) and Pubmed trying to identify the pathogen. The immunologist will use his or her own set of databases to understand pathogenesis and to identify therapeutics.

Response is a complex, highly branching process, requiring input and coordination across many disciplines, including medicine, public health, epidemiology, microbiology, genomics, and pharmacology. As mentioned above, there are currently hundreds of databases of biological knowledge available electronically across these fields. But the standards under which these data are recorded, searched and shared vary from one community to another. Even such fundamental entities as proteins and genes are burdened with multiple conflicting terminological and notational standards. Improved ability to share data will accelerate the pace at which pathogens are identified and treated. These data standards include such aspects as data format, terminology, and codes. Lack of agreement on these standards prevents the sharing of interoperable data, limiting data exchange and limiting the pace at which pathogens can be identified and characterized.

Semantic differences among research communities are such that collaboration is difficult and cumbersome. Clinicians, public health doctors, molecular biologists, first responders, and law enforcement speak different languages and have difficulty sharing data. Indeed, they operate on radically different scales, from molecules to individual people to populations, and may themselves be geographically dispersed.

The health care providers should be presented with appropriate patient information and medical knowledge at the point of clinical decision-making and records of clinical concepts and events in computable ways that can be accessed by the molecular biologist to further his or her own investigations. Controlled vocabularies are essential for computerized decision-support tools that will improve the quality of health care. Medical language must be recorded in standard ways so its meaning can be shared with other systems in a manner that is interoperable and computable. It is also essential to describe clinical concepts (problems, diagnosis, test results, and procedures) and laboratory concepts (characterization, identification, mechanism, and therapeutics) in an interoperable and coordinated manner. The lack of standardization, particularly of quantitative data and metadata, hinders interoperable use and requires a great deal of work to translate between databases and systems.

Semantically linked ontologies would allow laboratory researchers to search multiple databases with one query and access databases they would not normally think to access, would help link research done by other labs, make that research searchable by multiple sources, and help bridge the language gap between the molecular biologist and the clinician. Accelerating the development and deployment of biomedical IT will save lives, improve the quality of care, and maximize the efficiency of health care to the soldier.

Going back to the scenario described above, computational tools integrating diverse biological knowledge can accelerate the time to identifying and characterizing the unknown pathogen. There is no doubt that we face the danger of a biological weapons attack or emerging diseases. The new technologies that we propose to develop will ensure a faster response to intentional biological weapons attack, or the emergence of a new pathogen.

The problem of developing better tools supporting integrated research in biology is highly relevant to Government organizations such as the DoD, DHS, and the Intelligence Community. Given the changing threat landscape worldwide, our nation's forces could be highly vulnerable to biological threats. To meet these threats, research and development must address a range of issues that span basic research in molecular and cellular biology at one end, to epidemiology, public health, and countermeasures at the other.

The aim of integrating information on a massive scale, which necessarily entails the development of effective ontologies, is directly relevant to DoD's vision for the future. As expressed in JV2010 this includes: "Information superiority achieved through global, affordable, and timely access to reliable and secure information for worldwide decision-making and operations." This overall aim is part of the Department's *Defense Transformation Initiative,* and pertains directly to a problem that DoD calls *Horizontal Integration*.

Horizontal integration of information resources includes supporting the military's ongoing operational medical and healthcare infrastructure (one of the largest health-care systems in the country), as well time-critical mission functions in time of war or in the face of a terrorist attack. The DoD is also a partner with other agencies of the federal government in safeguarding civilian public health in response to an act of biological terrorism.

Protecting civilians and military personnel from bioterrorism involves long-term R&D, continual policy development, and emergency response planning. It also requires rapid, effective characterization of threats and timely response in the face of a potential attack. Knowledge systems technologies such as bio-ontologies that cover knowledge in basic biological research, medicine, and epidemiology are critical to these functions.

Current State of the Art

Addressing the problem of knowledge integration for biological threat response requires several different kinds of technologies, each of which is being pursued to some extent in the research communities and in some cases in industry. In this section we outline those technologies and their current state of the art.

Database Integration

One of the key components of a knowledge integration effort is the ability to integrate in some fashion the databases which contain much of the source knowledge; either through development of a common query language supported across the set of databases (*multidatabase* approach) or through integration of the set of databases into a single physical database system with a core shared schema (*warehouse* approach).

Two major efforts currently exist to address cross-database querying in the biomedical domain. The National Center for Biotechnology Information (NCBI) has developed the Entrez system[vi] which allows a user to enter a search to be run simultaneously against the databases represented within it (currently a small fraction of the available biological

databases). However, Entrez does not include any true integration of the accessible databases in that information in different databases is not explicitly linked in any way – rather the search queries each database individually, and the results are returned on an individual database basis. It provides a convenient interface for querying several databases simultaneously, but does not do any mapping to indicate where information in distinct databases might be identical or complementary.

The SRI Biowarehouse system is a bioinformatics database integration platform which loads various databases into a common database schema, such that distinct databases containing the same biological data types are coerced into the same table structure within the warehouse. In the current system, loaders handling the coercion of the individual databases to be integrated into the database schema must be written manually. The lack of automated tools supporting mapping of an external database structure to the common Biowarehouse schema limits the number of databases that can be integrated, and the speed with which they can be added. Perhaps more critically, while databases which are integrated into the warehouse share a common *schema,* they lack a common *vocabulary* such that querying across data derived from different sources is only marginally successful in many cases, despite the integration of these data sources within a common infrastructure.

In contrast, the Transparent Access to Multiple Bioinformatics Sources (TAMBIS[viiviii]) system provides transparent information integration and retrieval and filtering from multiple heterogenous biological information services by building a homogenizing layer on top of the different sources. This layer uses a mediator (information broker) and many source wrappers to create the illusion of one all-encompassing data source. It uses a rich, source-independent, ontology of molecular biology and bioinformatics to provide a unified conceptual level representation of its component resources. Local source concepts are mapped into this global ontology. The ontology provides a "language" for expressing complex queries over the domain, ranging over multiple diverse sources transparently. Though the TAMBIS ontology was designed specifically for the task of retrieval over bioinformatics resources, and may not be appropriate for other biological tasks, it stands as an example of the important role an ontology can play in addressing the information integration problem. It was developed manually, however, precluding rapid migration to new tasks or domains.

Knowledge Systems in Biological Network Analysis

Network analysis has become a common feature of biology in recent years. Although the complexity of biological systems has been appreciated for many years, high-throughput biotechnology has enormously increased the rate of information production. Because of this, there has been a resurgence of interest in complex systems modeling in the community that currently goes by the name of "systems biology."

The view of biological systems as networks of interacting elements spans a wide range of contexts. Genetic regulatory networks and cell signaling pathways are critical to understanding cellular function. Metabolic pathways are the key to understanding an

organism's internal functions, and at the community level one may view epidemiology and ecology as essentially network models.

Biological network analysis takes several distinct forms depending on its aims, progressing in one's understanding of the system of interest:

1. In **system characterization**, the objective is to describe the actors within a system and their interactions or dependencies based on analyses of various sources of evidence.
2. Given a basic picture of the connections in a system, **static** or **steady state** analyses can be performed. Many of these use graph theory to explore the topology of the network. Others, such as flux analysis, use basic conservation properties (e.g. mass-flux balance) to examine asymptotic or steady-state features such as the throughput of a metabolic network.
3. Finally, **dynamical modeling** aims to explore detailed function and predict the response of the system to stimulus. For some kinds of systems, such as those in cell signaling that process signals by exploiting system dynamics, it is nearly impossible to understand the system without a dynamical model.

The range of resources that systems biologists use in developing their understanding of biological systems can be very broad. For system characterization, they may explore structured data resources for functional genomics and proteomics, or may attempt to mine the literature. The GO and MeSH are commonly used.

System characterization is supported by analysis of function and interaction, where researchers rely on a host of resources in which the functional elements and dependencies of biological networks have been captured and preserved. These could be called pathway databases. One of the most famous is the *Kyoto Encyclopedia of Genes and Genomes* or KEGG. Another excellent example is the BioCyc[ix] collection maintained by SRI International.

Ontology development efforts to capture the essential knowledge elements of pathway representations are relatively recent and ongoing. Many of these are connected with specific data-producing enterprises like the Alliance for Cellular Signaling (AfCS[x]), or with analysis frameworks such as aMAZE[xi]. One recent effort to produce a common pathway ontology is BioPAX[xii].

Developing dynamical models requires numerical simulation, mapping the essential features of a biological network onto a mathematical model. Commonly this is to systems of differential equations, and the parameters that determine dynamical behavior must be acquired or inferred. The most visible community effort in this regard is the BioSpice[xiii] program sponsored by DARPA. This program's aims cover a very broad range of concerns including the development of simulation tools, the information system required to support them, and software architectures to integrate the entire enterprise of dynamical model construction.

To support this kind of activity, one must add a new area of knowledge representation to those already discussed, namely the dynamical modeling process itself. The most active community effort in this regard is the Systems Biology Markup Language (SBML[xiv]), itself a component of an earlier, more comprehensive effort, the Systems Biology Workbench (SBW[xv]). SMBL has developed a substantial life of its own due to the demand for interoperability among simulation tools in systems biology, and is supported by a number of simulation efforts in biology, most notably BioSpice.

The efforts described above are only a portion of the work underway in systems biology, which is becoming an extremely active field of research. Ontology development associated with this activity is substantial, and we have named only a few examples in our discussion. Nevertheless, the need for continued development of ontological resources and technologies, for example a comprehensive ontology to bridge all of the areas mentioned, is often acknowledged by this community.

Biomedical Ontologies

Ontologies in this context should be understood as formalizations of a domain of knowledge to facilitate communication between humans and computers. They consist of a collection of semantic information, typically sets of concepts and relations among them, which define a particular domain. They are related to informal resources such as thesauri and taxonomies, which adequately support human communication but do not provide the more formal and explicit concept definitions required by computer interaction. The challenge is to build ontologies that are useful and maintainable by humans, but also formal and interpretable by machines.

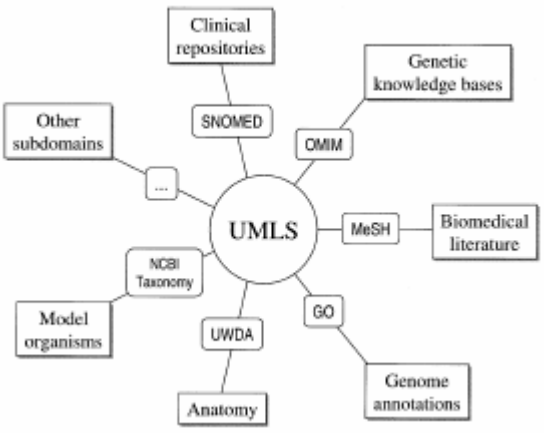
There currently exist a plethora of ontologies and taxonomies in the biomedical domain, and scores of database schemas and mark-up languages. Each is the result of a small group of scientists working without the benefit of knowledge-representation standards and knowledge modeling conventions. Scientists don't always know what ontologies even exist, what they might be good for, and how they relate to one another.

By far the most significant and successful of these is the GO[iii,iv] a large, hierarchical, ontology designed to allow data sharing between model organism databases by annotation of gene products with standard GO concepts (in this way it can also be viewed as a controlled vocabulary, augmented with semantics). It is divided into three branches, molecular function, biological processes, and cellular components. Genes are annotated with GO concepts, based on what is known about the function and location of the gene product (the protein). Each branch is also equipped with two semantic relations: is-a (subsumption) and has-part (composition). Although there are quite a few ontology development efforts in biology, the GO has evolved into a nexus for interrelating terms across genomes of widely divergent organisms, and is especially useful in systems characterization. GO has succeeded because it is usable by biologists and because it has served a pressing need – to support comparative genomics. Its success can be measured by the large number of model organism databases that have adopted it.

Other prominent ontologies include:

- The Digital Anatomist Foundational Model of Anatomy (FMA[xvi]) is a declarative domain ontology for human anatomy.
- NCICB's Enterprise Vocabulary Services(EVS[xvii]) for structured description of cancer data.
- SNOMED/CT[xviii] is the *de facto* standard for clinical terminology for both primary and secondary care and clinical trials.
- The Drug Ontology Project[xix] is a follow-on to the GALEN projects[xx] supported by the NHS, targeted an ontology and knowledge base for primary care prescribing.
- Biocyc uses an ontology to perform data integration through a central data warehouse.
- caBIG[xxi] (the Cancer Biomedical Informatics Grid) supports data and tool sharing related to the prevention and treatment of cancer.

There are also several efforts to bring these ontologies together into a common framework, notably the UMLS and the Open Biological Ontologies initiative (OBO[xxii]) OBO provides a single repository for open source ontologies in the biology domain, and requires that they share a common syntax, but does not include mappings or linkages among the ontologies that are stored there or connections to a common upper-level ontology.



The UMLS relates synonymous and similar terms via a common structure so as to tie together a variety of vocabularies[i]. The figure illustrates how the UMLS can serve as a link between not only vocabularies, but also the subdomains they represent. It can be used to collect the various terms used to name a concept, relationships among concepts, or concepts associated with a given category.

As such, the UMLS provides probably the most promising starting point for ontology and terminology integration. However, due to its requirement of maintaining "source transparency", i.e. representation of every relation asserted in a source vocabulary, it contains some inconsistencies and conflicts. Relations in different source ontologies cannot always be interpreted in the same way and terms from the source vocabularies may have some context-dependence in their interpretation. The juxtaposition of these within a single Metathesaurus in the UMLS inevitably introduces some incoherence in the resulting integration.

The state of the art in biomedical ontologies is one of proliferation, but with a wide range of quality and standards and a lack of integration. We need to improve existing

ontologies, to make them accessible and usable, and to enable integration among them. Even within a single ontology, and in particular for those ontologies developed by domain experts rather than knowledge modelers, there will be inconsistencies or gaps in coverage. For these resources to be important foundations for supporting bioinformatics work, and in particular to support sophisticated reasoning and query processing, these problems must be identified and resolved.

Ontological Technologies

The needs of biomedical applications have commonly driven knowledge systems technologies as developed in computer science, from database technology through expert system reasoning for diagnosis to semantic exchange environments. This continues to be the case today, and it is not surprising that two of the strongest general efforts in different knowledge systems areas, Stanford's Protégé ontology environment[xxiii] and the Ontology Web Language (OWL[xxiv]) for semantic exchange, were developed from a biomedical context.

Central to any interoperable ontologies is a language for their definition and exchange, and OWL is the current front-runner here. Built on the knowledge exchange standard RDF/RDFS, it provides a range of capabilities, from the definition of simple taxonomies, to constructs for explicitly defining relationships between concepts, novel constructs for describing properties and classes, and a rich set of necessary logical properties such as disjointness, cardinality and equality, enumerated classes, and more.

While OWL has just passed its first standardization process to become a W3C recommendation, several features have been omitted for simplicity and tractability which are essential for the definition of life science concepts (e.g. qualified cardinality constraints and complex role inclusion axioms). OWL emerged from the DAML+OIL language, originally developed within a prior DARPA DAML program[xxv].

An ontology specified in a language such as OWL is virtually impossible to support without some kind of logical reasoner to maintain an internally consistent structure. While several reasoners have been used in combination with OWL (e.g. F-logic[xxvi], Prolog and First Order logic theorem provers[xxvii]), description logics have received the most attention, since they have been designed specifically to reason over concept definitions found in ontologies. Highly efficient algorithms have been developed to make the task of checking consistency and subsumption between concepts tractable[xxviii] and the Description Logic Implementation Group (DIG[xxix]) has developed a standard interface to allow different Description Logic reasoners to be easily swapped in to an application.

Although we are advocating automated and semi-automated approaches to manage large ontologies, there will always be a need for interactive development environments analogous to those used for conventional software programming. The Gene Ontology Consortium has developed DAGEdit[xxx] for ontology authoring tailored to the GO with its limited number of relationships. The developers are currently adding OWL support, which may make it more generally applicable. Other ontology environments include the frame-based Generic KB editor developed for EcoCyc and NCICB's EVS.

However, the most widely used ontology-authoring environment is currently Protégé, which provides a plugin architecture for the addition of third-party functionality, a configurable "forms" interface for rapid authoring, a change management plugin "prompt", and support for collaborative authoring. Protégé was designed to conform to the frame-based Open Knowledge Base Connectivity standard[xxxi] that emerged from the DARPA High Performance Knowledge Base program. Protégé/OWL[xxxii] is a recent project to add OWL editing functionality to Protégé, which has successfully reconciled the frame-based view of Protégé with the logic axiom based view of OWL and description logic. Other collaborations between Manchester and Stanford Universities aim to build on Protégé/ OWL to develop further user-oriented OWL authoring tools.

Providing formal concept definitions in current ontology languages such as OWL is a non-trivial exercise that takes a substantial degree of training. Intermediate Representations (IRs) are domain specific ontology macro languages, which are much simpler than the description logic language itself. Multiple IRs were used in the GALEN-IN-USE project[xxxiii] to address the conflict between users with domain expertise and users with ontology expertise. Twenty thousand surgical procedure concept definitions were authored by surgeons in 3 days, compared to 3 months training to write in the underlying ontology language.

SNOMED/CT is a large description logic-based medical terminology supporting the authoring of 350,000 concept definitions. However, these tools are proprietary and so it is not possible to further comment on their functionality or the expressivity of the underlying description logic. They have no stated position on OWL or other ontology efforts.

Text and Natural Language Processing

Since terminological variation is highly problematic in the biomedical domain, and publications remain a primary vehicle for dissemination of biological knowledge, text processing must play an essential role in data integration and analysis.

There are several programs which aim to address aspects of text processing relevant to the needs of the proposed program. The ARDA AQUAINT[xxxiv] program is focused on natural language question answering, specifically addressing a scenario in which questions are asked in a focused topic area by a skilled, professional information analyst who is attempting to respond to larger, more complex information needs or requirements. There is currently no biology focus in this program; the focus is strictly on querying over text sources rather than over structured knowledge sources, but the basic scenario is highly relevant.

The National Business Center has organized a new program entitled "Research on English and Foreign Language EXploitation" (REFLEX[xxxv]). This program began September 2004 and is focused on information extraction from documents, for representation in some formal language or data structure. Tasks center on extraction of entities, relations, and events. The program also expressly incorporates extraction from

foreign language documents which is perhaps less relevant to the biology domain. As the program has barely started, it is unclear what will be accomplished that can be drawn on for bio-threat response but the research coming out of that program is potentially very useful.

There are important efforts underway in text mining in the biology domain. These are primarily focused on annotation of biological entities based on textual data. The BioCreAtIvE evaluation[xxxvi] addressed two main tasks: identification of gene or protein references in text, and assignment of GO annotations to proteins based on the information in a given document. The goal was to focus on tasks of relevance to biologists, and the results indicated that, in particular for the functional annotations, there is still significant room for improvement prior to adoption of the technologies by the biology community.

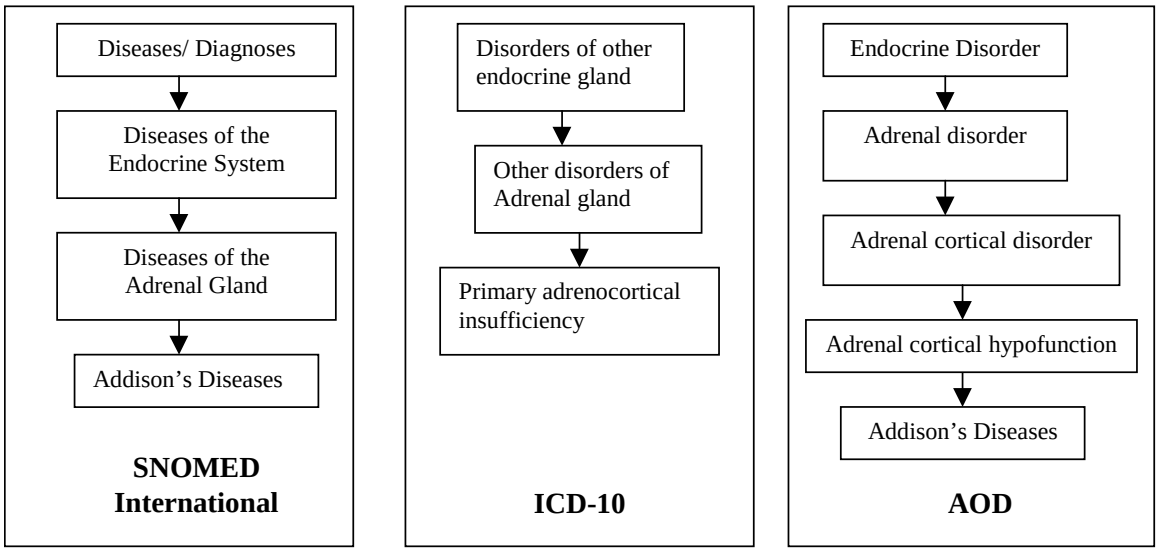
NIST sponsors an annual evaluation in text retrieval which now includes a genomics track (TREC Genomics[xxxvii]). In 2004 this included a task for *ad hoc* retrieval of documents from MEDLINE in response to a natural language query reflecting the information needs of biomedical researchers; and a categorization task consisting of two subtasks, a triage task aimed at identifying documents containing information relevant to annotation of a gene to a GO node, and an annotation task aimed at specifying which of the 3 branches of the GO the gene will be annotated to based on the information in the document. These two evaluations provide a forum for comparative assessment of techniques in text mining for the biology domain.

As an example of the challenges facing semantic interoperability stemming from terminology discrepancies and showing a need for representation of semantic relations, consider the following:

1. There are large disparities among the terms used and the codes generated by the various healthcare vocabularies. Table 1 illustrates the terms and resulting codes for the disease commonly referred to as “Addison’s Disease” in several current medical vocabularies. The problem is more acute than the flat table indicates: all of the vocabularies are hierarchical, and each of them has its own hierarchy structure. The figure below illustrates portions of the access paths to the concept of “Addison’s Disease” in three of the vocabularies. Each vocabulary makes different subdivisions of the domain, and the basis for the distinctions introduced is not made explicit.
2. There are no semantics associated with the terms in the vocabularies that would allow the recognition and reconciliation of different terms that have the same meaning. The shared meaning varies from “shallow” (e. g., near synonymy, as in “insufficiency” vs. “hypofunction” in the figure below) to “deep” (e. g., relations requiring more domain knowledge, such as “adrenal gland” is part of the “endocrine system”).

Vocabulary	Term	Code
UMLS	Adrenal Gland Diseases	C0001621
MeSH	Adrenal gland Diseases	D000307
AOD	Adrenal Disorder	0000005418
Read Codes	Disorder of the adrenal glands	C15z
SNOMED	Diseases of the adrenal glands	DB-70000

Table 1. Terms and codes used for "Addison's Disease"



Technology Gaps

Research is clearly continuing in each of the areas detailed above, and the communities dedicated to specific kinds of content continue to increase the availability of that content in a computer-accessible form. However, we have identified significant gaps specifically in available *technology*. Furthermore, while substantial progress in many of these areas is possible within a reasonable time horizon of development, they are also beyond any currently anticipated funding availability. It is thus within this gap that we identify the need for a future program.

Before calling out specific technological goals, we can generically identify at least the following kinds of technologies as being the most necessary, were they available:

- **Semantic Interoperability:** The need to support *ad hoc* queries points to what is by far the biggest goal, which is the support for interoperability of knowledge bases at the *semantic* level, that is, on the basis of the *meaning* of the information, rather than just the sheer ability to access shared records. The challenges of interoperability and integration will need to be addressed by the specification and verification of mappings between the knowledge elements (e.g. ontology nodes or vocabulary terms) in such a way that the semantics of terms are preserved. While

efforts here tend to focus on "top-down" strategies of terminological normalization and standards development across domains, such approaches will always be limited by the sheer size, number, and dynamics of change in existing databases, and the breadth and heterogeneity of the participating communities. So, in particular, we should be aiming at "bottom-up" technologies which allow novel knowledge resources to be "dropped into" an existing knowledge base, with the relevant semantic structures able to be aligned or linked together on an automatic, or at least semi-automatic, basis. This integration could focus at the level of term sets and controlled vocabularies, database schema, or full-up ontologies.

- **Supporting Inconsistency:** Such bottom-up integration will necessarily be imperfect. Thus techniques for representing and querying over integrated knowledge representations which accommodate incompleteness, inconsistency, unsoundness, and uncertainty, are necessary. At the least such issues must be recognized and represented, if not actually resolved.
- **Ontology Induction:** There is a huge need for what might be called "just-in-time" ontology construction, that is, to create perhaps smaller, perhaps lightweight, but novel ontological structures inductively from other sources such as database schema and texts, to be applied to specific *ad hoc* tasks. Machine learning can be employed to identify different names for the same biological entity, meronyms (e.g. parts of a complex), hyponyms (e.g. examples of kinases), and other semantic relations from text corpora.
- **Populating Existing Technologies:** Techniques to assist coercion of resources into an existing ontology are also important, such as the extraction of facts and relations from text and automated annotation or keyword extraction from text and structured data.
- **Natural Language Processing:** There are a range of NLP tasks which are relevant, including development of improved data- or knowledge-base query interfaces accommodating complex relational structures or cross-database querying. Terminology management and techniques to link linguistic with ontological structures are clearly of primary importance for enabling database interoperability. Ontologies can be used for automated expansion of keyterms in a query, for instance converting a query for "acetylcholine receptors" into one for ("acetylcholine receptors" or "muscarinic receptors" or "nicotinic receptors"). In the other direction, the knowledge embodied in an ontology can be used to help constrain interpretation of documents (e.g. word sense disambiguation) and to support extraction of entities and relations from text.
- **Provenance:** There are a range of issues surrounding the *provenance* of knowledge: if an ontology is used in a particular workflow, it must be possible to audit or trace what ontological knowledge contributed to particular outcomes, and then back through the well defined steps in the ontology construction methodology to determine the validity of that knowledge. Issues include versioning and revision, updating strategies, and annotations of quality and certification of data.
- **Generic Knowledge Systems Technologies:** Finally, in addition to the specialized needs within the bioinformatics community, there are also a number of issues in generic knowledge systems technologies where improvements can be

critical, including support for heterogeneous knowledge types (e.g. documents and images); more comprehensive methods for representation and reasoning; complementary views of knowledge bases (e.g. from the perspectives of a clinician vs. a genome researcher); support for multiple reasoning strategies (e.g. abductive, homological, and analogical); and more robust and efficient reasoners and inference strategies.

Multiple research communities are pursuing diverse technological agendas ranging from methodologies for building ontologies to procedures for involving a community in their maintenance, tools for authoring and updating, and tools to assist in consistency checking and consistency. Considering potential sources of Government funding, we note that the NSF funds small-scale computer science work on ontologies, but will not fund projects with "disease-specific" goals; the CDC BioSense program[xxxviii] has a mandate to integrate data for bio-surveillance, but has no research program in this area; while the NIH supports intramural research on ontologies (principally for cancer-related goals at NCI), the focus is primarily on human disease and less on basic research in microbiology, to understand the range of microorganisms and what can lead them to become pathogenic. NIST has supported work on national ontologies for manufacturing, but has only recently recognized biomedicine as an area of commercial interest; and finally standards bodies such as HL7 do not fund research.

We can therefore identify the following as a partial list of particular ontology technological goals necessary to advance the needs present in applications such as bio-threat response.

- **Combinatorial Algorithms and Order Theory:** Considered strictly as data objects, ontologies and taxonomies are rooted in a particular mathematical structure based on partially ordered sets (posets), similar to lattices. While posets and lattices are common in subsumption-based knowledge architectures such as object-oriented meta-models, their prominence in very large databases such as the GO are forcing new tasks such as navigation, categorization, and clustering in such ordered structures, and how multiple posets can be efficiently intersected and aligned. This more generally demands new mathematical concepts and combinatorial algorithms related to distance and level in lattices and posets.
- **Technologies for Ontology Tools:** Including algorithms for navigation, visualization, browsing support (e.g. alternative views) and version control.
- **Reasoners:** As noted above, while efficient reasoners are available, the biomedical domain is demanding in terms of both number and complexity of concept definitions. Work still needs to be done both to turn efficient reasoning components into effective reasoning components, and to package these reasoners to allow check-pointing of internal state, incremental classification, and debugging.

References

i . Bodenreider, O. [2004] The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, Vol. 32, Database issue.

ii. http://www.nlm.nih.gov/research/umls/

iii. http://www.geneontology.org

iv. Gene Ontology Consortium [2000]. “Gene Ontology: A Tool For the Unification of Biology", *Nature Genetics*, v. 25:1, pp. 25-29.

v. http://www.gideononline.com/

vi. http://www.ncbi.nlm.nih.gov/Entrez

vii. http://imgproj.cs.man.ac.uk/tambis/

viii . Goble, C.A., R. Stevens, G. Ng, S. Bechhofer, N.W. Paton, P.G. Baker, M. Peim, A. Brass. [2001] Transparent access to multiple bioinformatics information sources. IBM Systems Journal, Vol. 40, No. 2.

ix. http://www.biocyc.org

x. http://www.signaling-gateway.org/

xi. http://www.amaze.ulb.ac.be/

xii. http://www.biopax.org/

xiii. https://community.biospice.org/

xiv. http://sbml.org/index.psp

xv. http://sbw.sourceforge.net/

xvi. http://sig.biostr.washington.edu/projects/fm/index.html

xvii. http://ncicb.nci.nih.gov/core/EVS

xviii. http://www.snomed.org

xix. http://www.cs.man.ac.uk/ mig/projects/old/drugontology/

xx. http://www.cs.man.ac.uk/mig/projects/old/galen/index.html

xxi. https://cabig.nci.nih.gov/

xxii. http://obo.sourceforge.net/

xxiii. http://protégé.stanford.edu

xxiv. http://www.w3.org/2004/OWL

xxv. http://www.daml.org/

xxvi. http://www.cs.umbc.edu/~hchen4/fowl

xxvii. http://www.cs.man.ac.uk/~riazanoa/Vampire

xxviii. Horrocks, I., U. Sattler, and S. Tobies. [2000] “Practical reasoning for very expressive description logics”. *Logic Journal of the IGPL*, 8(3):239-263.

xxix. http://dl.kr.org/dig/

xxx. http://geneontology.sourceforge.net

xxxi. http://www.ai.sri.com/~okbc/

xxxii. http://protege.stanford.edu/plugins/owl/

xxxiii. http://www.cs.man.ac.uk/mig/projects/old/giu/index.html

xxxiv. http://www.ic-arda.org/InfoExploit/aquaint/

xxxv. http://www.nbc.gov/reflex.html

xxxvi. http://www.mitre.org/public/biocreative/

xxxvii. http://medir.ohsu.edu/~genomics/

xxxviii. http://www.cdc.gov/phin/Webinars/BioSense.htm