Goals and Outcomes Report for the Infectious Disease Ontology Workshop and Meeting
Cold Spring Harbor Laboratory, September 19 – 21, 2007
Organizers: Lindsay Cowell and Barry Smith

Biomedical research is in a state where sound ontologies are desperately needed to enable integration, exchange and reuse of data. High-throughput technologies have created new data management problems, and researchers are under ever more pressure from funding agencies to share data, meaning that data resources must be interoperable. Ontologies are useful for much more than data management, however, and can be used to support sophisticated computational analyses of data and to facilitate human understanding of data which crosses disciplinary boundaries. They also support training and education and cross-domain collaboration. Many areas within biomedicine, however, still have no ontology coverage at all, making it impossible for researchers in these areas to share in and contribute to the achievement of such benefits. Furthermore, the vast majority of those biomedical ontologies which do exist are either so full of errors or gaps that they are unusable, or they are useful only for the particular purposes of the particular group of individuals who created them. This is largely due to a lack of awareness among biomedical researchers of good ontology development practices and a lack of trained researchers in biomedicine with ontology expertise.

To contribute to the solution of these problems, we planned a three-day training event focused on the development of an ontology of infectious disease. We selected infectious disease as a specific target domain because infectious disease research presents specific data management and analysis problems that can be greatly alleviated by the use of ontology-based methods.

The specific goals of the three-day training event were to:

1. raise awareness within the community of infectious disease researchers about the benefits of ontology to their work
2. raise awareness in the community of ontology developers about the specific needs for ontology in infectious disease research
3. train infectious disease researchers in ontology development best practices
4. create a draft infectious disease ontology
5. engage a community of researchers in the continued development, testing, and use of the infectious disease ontology

Workshop and Meeting Agenda

To accomplish the above goals, the three-day training event was organized as a two-day training workshop and one-day public meeting, where the focus of the workshop was on training and ontology development and the focus of the public meeting was on community outreach and planning future directions.

The first day of the workshop consisted primarily of presentations by the organizers providing relevant introductory material. The schedule was as follows:

- 9:00am to 12:30pm - Background Session 1: Introduction to immunology and infectious disease, designed to give an overview of the domain, especially for informatician and ontologist participants.
    - Components of the immune system
    - Induction of an immune response
    - Types of pathogens and types of immune responses
    - Steps in pathogenesis
    - Pathogen tactics for evading the immune response
- 2:00pm to 5:30pm - Background Session 2: Introduction to the principles of ontology development.
    - What is an ontology?
    - Open Biomedical Ontologies (OBO)
    - Basic Formal Ontology
    - The OBO Foundry methodology
    - Principles of development
- 7:00pm to 9:00pm - Background Session 3: Ontologies – the Good, the Bad and the Ugly.
    - What are ontologies useful for in biomedicine?
    - Success stories
    - Overview of existing resources
    - How do we begin to create an ontology of the infectious disease domain within the OBO Foundry framework?

The second day of the workshop opened with examination of a draft infectious disease ontology and continued with development of the ontology over the course of the day, providing workshop attendees the opportunity to apply what they had learned the day before and to gain experience with ontology development. This group ontology development session utilized a small draft infectious disease ontology prepared by Dr. Cowell and a member of her research group, Anne Liebermann, beforehand. The session provided the opportunity for workshop attendees to contribute directly to the infectious disease ontology by filling in gaps and pointing to errors and areas of unclarity in this draft ontology, thereby increasing the attendees' interest in remaining involved in the ontology development effort after the workshop's conclusion.

The draft infectious disease ontology prepared for the workshop contained 129 terms organized into four types taken from the Basic Formal Ontology upper-level ontology. The 129 terms were drawn from both immunology and infectious disease research. They included 58 terms for processes (e.g. transmission), 5 terms for independent continuants (e.g. vaccine), 30 terms for qualities (e.g. immuno-compromised), and 36 terms for roles (e.g. host). The number of terms for independent continuants is small because many of the independent continuants relevant to infectious disease research will be included in other ontologies. For example, terms for anatomical entities will be included in the Foundational Model of Anatomy.

During the second day of the workshop, the draft infectious disease ontology was revised, the scope of the ontology was discussed, and a mid-level organization was proposed. The

draft ontology was revised by deleting and adding terms and by creating definitions for some of the terms. It was decided that the resultant Infectious Disease Ontology (IDO) should be a disease-neutral core, in the sense that it should cover those entities – such as host and pathogen – which are relevant to all infectious diseases, and that disease-specific extensions would be created from this core by working with corresponding specialists. With IDO as a core ontology, the relationships between terms in different disease-specific ontologies will be determined by the relationships of these terms to terms in IDO, thereby ensuring interoperability between the disease-specific ontologies and comparability of the data annotated using terms drawn from the latter

The specific scope outlined for this IDO core ontology include the following domains:

- pathogen, vector, host, and interactions between them;
- disease at the level of the individual and at the level of populations;
- biology of infectious disease;
- clinical diagnosis, treatment, care, and prevention of infectious disease.

The importance of identifying areas within these four domains that overlap with existing ontologies and of importing terms from those ontologies was emphasized. For example, many of the processes relevant to the basic biology of disease at the level of the individual are included in the Gene Ontology Biological Process ontology (GO BP); thus these terms will be imported from GO BP and consistency with GO BP will be maintained in the future. It was further emphasized that IDO and its projected disease-specific extensions should be created with a modular structure that will allow easy migration of terms to ontologies that may be created in the future and promote division of labor in development and maintenance in the future. For example, a clinical care ontology might include terms relevant to the diagnosis and treatment of disease, leaving only those terms specific to the diagnosis and treatment of infectious disease for inclusion in IDO.

The training workshop concluded with a discussion of potential uses of IDO in annotating the data being assembled by specific groups. Agreements were reached with these groups concerning cooperative development of IDO; these agreements are outlined below.

The public event on the third day was organized into three sessions:

- Topics in Infectious Disease Research
- Ontologies and their Application to Infectious Disease
- Ontology and the Future of Infectious Disease Research.

The first session consisted of two presentations, one by Dr. Stefan Kaufmann entitled *Global Threats Need Global Research Efforts: The Example of Tuberculosis* and one by Dr. Ronald Veazey entitled *Utility of Nonhuman Primates for Examining Transmission and Pathogenesis of HIV Infection*. Dr. Kaufmann is Director of the Department of Immunology at the Max Planck Institute for Infection Biology in Berlin and Professor of Microbiology and Immunology at the Charité, Humboldt University, Berlin. Dr. Kaufmann has made significant contributions towards our understanding of immunity to bacterial pathogens resulting in advances in the development of a tuberculosis vaccine. His research has focused specifically on the following areas: tuberculosis gene expression

at different sites within the lung, biomarkers for discrimination between latent infection and active disease, and the role of T cells, especially regulatory T cells, in the immune response to tuberculosis. He outlined the international efforts to coordinate research, treatment and prevention of infectious disease. Dr. Veazey is Professor of Pathology and Chair of the Division of Comparative Pathology for the Tulane National Primate Research Center. His research on SIV in non-human primates has made significant contributions to our understanding of HIV transmissibility and pathogenesis. His seminal work demonstrating that early events in viral transmission and replication occur primarily in the intestinal and vaginal mucosa has led to a shift in focus in HIV research from blood to mucosal immunity and has led to significant efforts to develop topical microbicides. His talk provided an overview of infectious disease research in non-human organisms and of the relevance of this research to the human case.

The second session consisted of three presentations. The first, *Surgical Endocarditis: Bringing Ontology into the Operating Room*, was given by Dr. Steve Gordon, Chairman of the Department of Infectious Diseases in the Cleveland Clinic's Division of Medicine. Dr. Gordon has a long-term interest in the prevention and treatment of transplant infectious diseases, cardiothoracic infections, and healthcare-associated infections. His work embraces the development and evaluation of innovative infectious disease diagnostics, including the planned development of an infectious disease ontology covering infections problematic in surgical contexts. His talk focused *inter alia* on the way in which ontology-based technology can help in assembling comparisons of outcomes data useful to the operations of surgical enterprises. The second presentation was given by Dr. Michael Ashburner on *Ontologies for Biomedicine: The GO Experience*. Dr. Ashburner is Professor of Biology in the Department of Genetics at the University of Cambridge and a Fellow of the Royal Society. He has contributed significantly to the field of biomedical ontology, primarily through his leadership role in both the Gene Ontology Consortium and the Open Biological Ontologies Project. The session concluded with a presentation by Dr. Lynn Schriml describing the Genomic Metadata for Infectious Agents (GEMINA) project. GEMINA is a web-based system designed to identify infectious pathogens and their representative genomic sequences through selection of associated epidemiology metadata (http://gemina.tigr.org/cgi-bin/MakeFrontPages.cgi). It relies on an ontology, whose content overlaps significantly with that of IDO and its projected extensions. The GEMINA ontology was created through the merger of multiple ontologies created with differing organizing principles and varying degrees of formal rigor. Thus, the GEMINA ontology does not serve the purposes we intend for IDO. We will, however, coordinate future efforts with Dr. Schriml in hopes of evolving towards one common ontology.

The final session of the meeting was a panel session with four participants: the meeting organizers, Dr. Richard Scheuermann, and Dr. Lincoln Stein. Dr. Cowell opened the session with a summary of the two-day workshop and of its results. Her presentation was followed by presentations from Drs. Scheuermann and Stein describing the BioHealthBase and Reactome information resources, respectively. Dr. Scheuermann is Professor of Pathology at the University of Texas Southwestern Medical Center. He is Principal Investigator for both the ImmPort and BioHealthBase information resources

and a powerful protagonist of ontology-based technology as a foundation for the future of information-based biomedical research. ImmPort is a system for the long-term, sustainable archival of all research data generated by the ~1500 investigators funded by the NIAID Division of Allergy, Immunology and Transplantation. BioHealthBase integrates genome sequence, functional genomic and related data critical to scientific research in the development of vaccines, therapeutics, and diagnostics for five types of priority pathogens. Dr. Stein is a researcher at the Cold Spring Harbor Laboratory and Principal Investigator of the Reactome information resource, a curated resource of core pathways and reactions in human biology, including pathways and reactions of great significance to the ontology of infection and immunology.

Dr. Smith closed the panel session by leading a discussion about future directions for the infectious disease ontology, and of the advantages and disadvantages of the core-and-extensions model of development we have selected for IDO in the future. Dr. Christos Louis, leader of the Insect Molecular Genetics Group in Crete and a specialist in the study of vector-borne diseases, offered strong arguments in favor of this model, based on the fact that the disease-neutral core can be completed in a relatively short space of time, and thus already serve to ensure a large degree of terminological coordination across multiple groups from a very early stage. This discussion included confirmations of commitments from eight different research groups to collaborate on various aspects of the core infectious disease ontology and disease-specific extensions to it as well as plans for a future meeting. These activities are described in more detail below under Future Directions.

Workshop and Meeting Attendance

The initial two-day workshop was attended by 28 researchers from differing career stages and a variety of backgrounds. Six of the attendees are graduate students; nine are informaticians or biologists employed as research scientists developing relevant information resources (such as BioHealthBase described above); five are postdoctoral research fellows; eight are faculty researchers. Six of the attendees are philosopher ontologists; eight are biologists from either a microbiology or immunology background; ten have an interdisciplinary background combining biology with database curation and management; four have a computational or medical informatics background and are working in infectious disease-relevant areas (e.g. disease surveillance, clinical decision support).

In addition to the 28 individuals described above, the one-day meeting was attended by four researchers from GEMINA, DUAL Knowledge Systems, Novartis, The Jackson Laboratory, and Digital Infuzion.

Twenty-six of the 28 workshop attendees completed evaluations of the two-day workshop. The evaluation consisted of six questions with scored answers and six questions soliciting comments (see attached). Scored answers can range from 1 (poor, definitely change) to 5 (exceptional, don't change at all). The average scores for all 6 scored answers range from 3.73 to 4.08 indicating an overall high level of satisfaction with the workshop.

Future Directions

The success of the three-day training event is most evidenced by the participants' commitment to continued involvement in development of the infectious disease ontology. We have commitments from two research groups to develop specific branches of the infectious disease ontology. Dr. Richard Scheuermann's research group at University of Texas Southwestern Medical Center has agreed to contribute the terms referring to pathogen-relevant entities, and Dr. Yonggun He's research group at the University of Michigan has agreed to contribute the terms referring to vaccine-relevant entities.

In addition, six research groups have agreed to develop disease-specific extensions of the infectious disease ontology. Initial development of these disease-specific ontologies will serve to evaluate and improve the infectious disease ontology, and their subsequent development will establish them as information resources. Dr. Christos Louis' research group will develop an IDO-based ontology of vector-borne diseases with a focus on malaria; Dr. Saul Lozano-Fuentes' group at Colorado State University will develop an ontology of Dengue Fever; Dr. Lindsay Cowell's research group at Duke University Medical Center will develop an ontology of tuberculosis; Dr. Sivaram Arabandi's group at the Cleveland Clinic will develop an ontology for infective endocarditis; Dr. Stuart Sealfon's group at Mt. Sinai Medical Center will develop an ontology for influenza; and Dr. Younggun He's research group will develop an ontology for brucilosis.

To facilitate continued interactions among workshop attendees and to provide a forum through which additional researchers can join the collaboration, we have established a wiki and email list for the infectious disease ontology. The wiki is maintained at http://www.bioontology.org/wiki/index.php/Infectious_Disease_Ontology. To subscribe to the email list, visit https://lists.duke.edu/sympa/. To email the list, write to ido@duke.edu.

Our current focus is to continue development of IDO and to publish a paper describing it as soon as there is a stable first release. This paper will further publicize the ontology encouraging other researchers to join the collaboration. At the same time, we will work to obtain funding to support curation of the infectious disease ontology and the development of computational applications to demonstrate its utility. Drs. Smith and Cowell have already planned the development of two applications of the infectious disease ontology, one in collaboration with Dr. Carol Dukes-Hamilton and the other in collaboration with Dr. Vance Fowler, both clinical researchers at Duke University Medical Center. In collaboration with Dr. Dukes-Hamilton, we have plans to develop a tuberculosis clinical decision support system based on the infectious disease ontology. In collaboration with Dr. Fowler, we have plans to develop an ontology-based approach to the identification of host susceptibility genes associated with *Staphylococcus aureus* infection.

Conclusion

In the last two years, there has been a surge of interest in ontology within the biomedical research community. Despite this increase of interest, many areas of biomedicine have

no ontology coverage, and biologists do not yet have the necessary knowledge to develop sound ontologies. Thus, the goals of this workshop were to provide a draft infectious disease ontology and ontology training for infectious disease researchers thereby laying a solid foundation upon which the application of ontology to the study of infectious diseases can grow rapidly. The workshop succeeded in meeting these goals. Researchers with backgrounds in immunology, infectious disease, informatics, and ontology were recruited to the workshop. Responses to the workshop evaluations indicate that the workshop succeeded in fulfilling its training mission. A draft infectious disease ontology was produced, and, most importantly, a large number of attendees with varying areas of expertise are enthusiastic about this ontology and have agreed to participate in its continued development and its application. The attendees agreed that a follow-up meeting should be held in 2008 to further consolidate the results of our training efforts and thus to continue growing the community of involved researchers.