Editorial Comments

Biosurveillance, Classification, and Semantic Health Technologies

CHRISTOPHER G. CHUTE, MD, DrPH


The emergence of the World Wide Web in the early 1990s catapulted the Internet from a curiosity for computer scientists, engineers, and experimenters testing data interchange technologies into a vast, dynamic, and unprecedented information resource that has fundamentally transformed science, society, and human communications. The implications of this intellectual paradigm shift are still being explored, and the full impact of this dramatic transformation, which instantly makes available massive amounts of information, has not yet become fully appreciated. Nevertheless, the emerging dependency of the health sciences on increasingly practical semantic technologies to organize and leverage these vast information resources is now unquestioned. Ranging from the interpretive functionality of the human genome to public health surveillance among nations and continents, the full spectrum of health science is being fundamentally transformed by the application of basic science informatics in the domains of meanings, ontologies, and natural language processing; these have matured into critical technologies and unquestoned adjuncts of the emerging big-science transformation of biomedicine and healthcare.

In this issue of the Journal, two articles illustrate some of these dependencies upon semantic health technologies. The first employs sophisticated dictionary lookup techniques to classify news and specialist articles about disease outbreaks as an adjunct to outbreak detection and severity measurement. The second posits a sophisticated scaling of outbreak severity based not only on disease metrics but also on sociological and governmental reactions in the face of mild to severe epidemics.

The article by Freifeld and colleagues describes classification and display software that systematically collects data from outbreak-alert distribution lists and general news media in order to classify disease and location. Once classified, the information is visually displayed on a Web-rendered world map with various capabilities for expanding or contracting time windows and locations. The classification engine itself involves a multistage parsing, part of which includes fast dictionary lookups against word-level N-grams using cascading hashes of keywords, locations, and organisms. Additionally, this dictionary has limited ontologic capabilities, such as “containment,” where it will explicitly recognize that the city of Boston is located in the state of Massachusetts. There are significant technical limitations, readily acknowledged by the authors, such as an undue reliance on exact phrase matching. On the other hand, this brute force dictionary lookup methodology has the advantage of generating results that are highly effective, and it illustrates the utility of such Internet surveillance for identifying outbreaks and their locations and for depicting their apparent intensity.

The work by Wilson et al. focuses on the severity of social disruption for biological outbreaks among humans or animals by examining “the indications and warnings (or markers) of social disruption used in tandem with explicit reports of disease.” The authors’ emphasis is on defining stages of outbreak severity, which include (adapted from Table 1): 0) Environmental conditions favorable to an outbreak, 1) Localized biological event, 2) Multi-focal biological event, 3) Severe social and medical infrastructure strain, 4) Social collapse, and P) Preparatory posture.

However, in their discussion, the authors explicitly raise the obvious strategy of Internet technologies as possible “harvesting engines” to capture information relevant to the definitional criteria for biological-outbreak severity metrics. The examples they provide on Rift Valley Fever, Venezuelan Equine Encephalitis, and SARS were all humanly annotated and curated. Nevertheless, the prospect of applying algorithmic methods to “harvest” the requisite information from broadly based, multilingual news media on the Internet to 1) identify emerging outbreaks and 2) assign initial outbreak severity or track the escalation of severity over time is the point relevant to this editorial.

While these manuscripts differ dramatically in their scope—the first focusing on disambiguating disease and location of outbreaks, the second spanning similar biological characteristics but also including sociological and governmental responses in the face of public health threats—they are similar in their existing or implied dependence on semantic health technologies. As biology and medicine evolve toward a big-science paradigm, analogous to the evolution experienced by physics early in the 20th century or astronomy in the latter half of the century, an obvious question for the authors of these two manuscripts and their underpinning infrastructures is whether any data sharing or interoperability can occur among their described systems—present or
future. Efforts to achieve an efficient public health infrastructure must obviously invoke Semantic Web principles to establish shared terminologies, ontologies, and knowledge metadata. It would be intriguing to establish whether the disease ontologies across these systems have a machinable overlap (it is assumed that they have an implicit semantic overlap). Assuming they do not, how can standard ontologies of disease outbreaks evolve? In a similar vein, the systems described in these articles are explicitly or implicitly dependent upon Natural Language Processing (NLP) of news and outbreak reports. To what extent do they share dictionary lookup technologies or more sophisticated language processing capabilities? While these questions are rhetorical in the present circumstance, they increasingly define the research agenda and infrastructure creation needed to achieve the visions captured by these articles.

Public health is not alone in its evolution toward big science. Basic biology, anchored on the human genome, has witnessed the emergence of the Gene Ontology as a nascent interlingua to describe biological functions associated with genes and gene products. Similarly, NLP techniques are harvesting the biomedical literature in an effort to build more coherent knowledge atop the computationally fragile collections of quaint text. In clinical medicine, NLP and data normalization form the bedrock of emerging clinical data warehouses, whose purpose explicitly is to improve the quality of care for existing patients and aid the discovery of new knowledge. Transcending these arenas is the Genotype–Phenotype Holy Grail, now manifest by many initiatives, including the NHGRI Genome-wide Association Studies (GWAS) linked with EMR-derived phenotype definitions; this translational science is even more dependent on semantic health technologies for generalizable success.

Among projects that seek to provide unity across this spectrum of genomic to phenotypic characterization, with application in basic science, clinical medicine, and most assuredly public health, is the next generation International Classification of Diseases (ICD) in early development by the World Health Organization (WHO). In my capacity as Chair of the ICD Revision Steering Committee, I can report that, while promising to be built on robust ontologic principles, with linkage to underpinning terminologies, the resulting high-level rubrics or classification spaces of the new ICD may be candidates for the kind of infrastructure that the two articles in today’s Journal might effectively leverage—at least with respect to disease. Corresponding work is ongoing with the new SNOMED (now known as the International Health Terminology—IHT), the Gene Ontology, and emergent products from the Open Biomedical Ontologies; all potentially coordinated by the newly established National Center for Biomedical Ontologies based out of Stanford.

To make NLP and efficient dictionary lookup techniques practical, a robust thesaurus of standard terminologies and classifications must evolve to include multilingual synonymy as well as phrase-level synonyms within languages. The challenge of creating such an open-source, fundamental linguistic resource for basic biological science, clinical medicine, and public health is slowly being recognized as the sine qua non for harvesting information efficiently and fulfilling the promise of biology and medicine by moving toward genuine big-science integration.

References