In: Baker CJO, Cheung K-H, editors. Semantic Web: Revolutionizing knowledge discovery in the life sciences: Springer; 2007. p. 101-119.

Chapter 5

CLINICAL ONTOLOGIES FOR DISCOVERY APPLICATIONS

Yves A. Lussier[1,2] and Olivier Bodenreider[3]

[1] Section of Genetic Medicine, The University of Chicago,USA; [2] Department of Biomedical Informatics and College of Physicians and Surgeons, Columbia University, USA; [3] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract: The recent achievements in the Human Genome Project have made possible a high-throughput "systems approach" for accelerating bioinformatics research. In addition, the NIH Whole Genome Association Studies will soon supply abundant clinical data annotated to clinical ontologies for mining. The elucidation of the molecular underpinnings of human diseases will require the use of genomic and ontology-anchored clinical databases. The objective of this chapter is to provide the background required to conduct biological discovery research with clinical ontologies. We first provide a description of the complexity of clinical information and the main characteristics of various clinical ontologies. The second section illustrates several methods used to integrate clinical ontologies and therefore databases annotated with heterogeneous standards. Finally the third section reviews a few genome-wide studies that leverage clinical ontologies. We conclude with the future opportunities and challenges offered by the Semantic Web and clinical ontologies for clinical data integration and mining. Discovery research faces the challenge of generating novel tools to help collect, access, integrate, organize and manage clinical information and enable genome wide analyses to associate phenotypic information with genomic data at different scales of biology. Collaborations between bioinformaticians and clinical informaticians are poised to leverage the Semantic Web.

Key words: Clinical Terminology, Clinical Ontology, Clinical Phenotypes, Discovery, Phenomics.

1. INTRODUCTION

Achievements in the Human Genome Project have made possible for a high-throughput "systems approach" to understand, prevent and treat human diseases. While the platform of molecular networks, especially gene profiling under homeostatic or disease conditions has been intensively explored as a gateway to "systems medicine," this approach to analyzing genomic data is often complicated by genetic heterogeneity and the lack of cellular, tissue, organ, anatomical or environmental context to accurately interpret the gene functions which are highly context-dependent. Further, as mutations in different genes may yield identical or related phenotypes, a molecular characterization solely based on genes may neglect important relationships between molecularly distinct diseases at the level of phenotype. While altered phenotypes are among the most reliable manifestations of altered gene functions that can be observed, described, and quantified, research using systematic analysis of phenotype relationships to study human biology is still in its infancy [1]. In addition, the advent of large scale genetic databases together with the NIH Whole Genome Association Studies have intensified the need for high-throughput discovery technologies to efficiently manage, access, integrate, and reuse the wealth of phenotypic and genomic data.

As we will describe in this chapter, Clinical Ontologies and related tools offer a unique opportunity to organize and access well-networked and integrated clinical phenotypes from otherwise heterogeneous information sources.

1.1 Complexity of Representation of Clinical Information

The issue of complexity of phenotypic information and knowledge representation includes (i) definition, (ii) composition, (iii) scale, and (iv) context. Clinical phenotypes are sometimes ambiguously defined. Mahner has found at least five different definitions of phenotypes in the literature [2]. Clinical Ontologies represent clinical phenotypes, diseases, syndromes and many other clinical elements such as medications and personal habits (e.g. smoking), which are considered "environmental conditions" in biological communities.

1.1.1 Ontologies and Terminologies

Ontologies and their associated systems [3-7] are robust architectures designed for knowledge representation of concepts and the relations among

them in a formal language (often frames or description logics). They have been widely used in biology and medicine [8-11]. However, few phenotypic terminologies satisfy these criteria [12]. Obstacles in modeling phenotypic knowledge in a formal ontology involve the difficulties and costs of (i) achieving consensus regarding the definition of phenotypic entities, and (ii) enumerating the *context features* and the background knowledge required to ascribe meaning to a specific phenotypic entity[13-15]. In this chapter, we adhere to a looser definition of Clinical Ontology, which also includes well-organized–but not always formally represented–clinical classifications, nomenclatures and terminologies.

1.1.2 Compositional Clinical Phenotypes

First we will provide examples of the compositional nature of clinical phenotypes, followed by the ambiguity that can arise from different information models representing these phenotypes. Clinical phenotypes are highly compositional in nature [14, 16-19], one can refine a phenotypic description with additional modifiers. For example, the concept right tibial dysplasia can be represented by associating the following components: {Regional Anatomy: Laterality: "right"} and {Systemic Anatomy: Bone: "tibia"}, characterizing an anatomical entity, which can be further modified by {Abnormal Anatomical Structure: Morphology: "dysplasia"}.

Information models help delineate which representation styles are used to store and query clinical phenotypes. When components of a composite phenotypic concept are implemented in a database schema, implicit knowledge about the composite clinical phenotype is buried in the information model. For example, "right tibial dysplasia" can be coded as a single field in a broad accident database, using a pre-coordinated term. In contrast, in order to support detailed queries with respect to anatomy and morphology, the same concept can be decomposed into several fields (possibly located in different tables) in the clinical information system of an orthopedic surgery department. While the information stored may be equivalent in both cases, the split terminological components of the overall concept can only be construed as equivalent to the whole by post-coordinating (i.e., reassembling) the overall concept using metadata often implicitly buried in the local information model [20].

1.1.3 Context of Clinical Phenotype usage

The context in which a clinical phenotype is stated is very important to its pertinent reuse. The meaning of a term varies with context in normal language, but context must be represented explicitly if one is to

meaningfully organize related phenotypic data, collected under diverse conditions or from distinct databases. For example, the views of different professions using a specific term may carry some implicit knowledge since the nature of the source database may not be associated with the concept. For example, the term "mole" found in a dermatology database does not carry the same meaning as in a gynecology database. While in dermatology, "mole" refers to a skin lesion, the "mole" phenotype in gynecology describes an intrauterine tumor [21]. Similarly, the context of the experimental conditions, the organism under study, and its stage of development may also significantly modify the meaning of a phenotype.

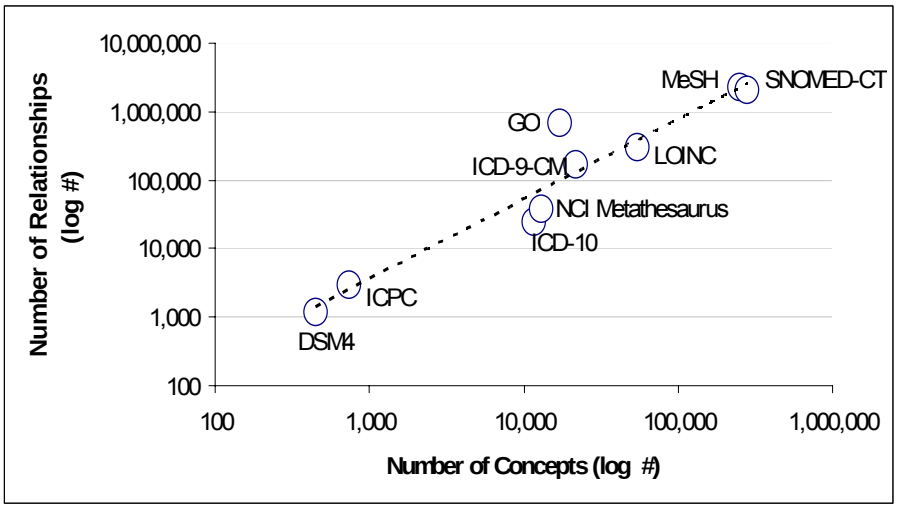

Figure 5-1. Quantitative Comparison of the Content of Clinical Ontologies

1.2 Clinical Ontologies, Terminologies, Classifications and Nomenclatures

This section will summarize the properties of clinical ontologies that are well known and used by different clinical communities to annotate datasets. Figure 5-1 provides an overview of the number of concepts and relationships in each ontology. The linear relationships between the axes of Figure 5-1 imply that relationships "*R*" in clinical ontologies are increasing as a power function of the number of concepts "*C*" (e.g. $R=C^n$), where "n" can be calculated from the figure. Table 5-1 provides the details on the clinical entities covered by these ontologies.

Table 5-1. Content coverage of distinct ontologies according to the scale of biology and scientific field. Legend: ●= biological scale covered, ○= biological scale partially covered, "empty box"= biological scale not covered

Scale of Human Biology	Clinical Ontologies					Scientific Fields
	ICD	MeSH	SNOMED	UMLS	NCI Metathesaurus	
Clinical Proteins		○	○	○	○	Proteome, interactome, Structural Genomics, Gene product pathways
Clinical Gene Functions		○	○	●	●	
Cell Morphologies		●	●	●	●	Histology, Pathology
Cell types		○	●	●	●	
Tissues, Morphologies		○	●	●	●	
Organs		●	●	●	●	Clinical Genomics, Pharmacogenomics
Systems		●	●	●	●	
Diseases, Syndromes, Populations	○	○	●	●	●	Medicine, Nursing, Public Health

1.2.1 Properties of Clinical Ontologies

Some ontologies are more convenient to compute with, due to superior design. Table 5-2 summarizes the different properties of each clinical terminology. Cimino proposed the following list of properties used in Table 5-2 to summarize the computability of clinical ontologies [22, 23]:

- *Concept-Oriented:* the preferable unit of symbolic processing is the concept.
- *Formal semantic definition*: the semantic definition of concepts in an ontology as defined in Section 5-1.1.1.
- *Concept permanence:* the meaning of a concept should not change over time and obsolete concepts are retired, not deleted.
- *Nonredundancy*: the definition of a concept should be unique.
- *Nonambiguity*: distinct concepts should not share the same terminology or code.
- *Relationships* between concepts differentiate expressiveness of ontologies:
  - *Monohierarchy (Tree):* each concept has only one parent.
  - *Polyhierarchy* : Concepts may have more than one parent.
  - *Directed Acyclic Graph (DAG):* no cycles are allowed in the graph.

Table 5-2. Properties of Biomedical Ontologies
legend: ●= property provided, ○= property partially provided, "empty box"= property not provided

Properties of the Ontology		Clinical Ontologies					
Class	Subclass	ICD-9, ICD-10	ICD-9-CM	MeSH	SNOMED CT	UMLS	NCI Metathesaurus
Architecture	Concept-Oriented	●	○	○	●	●	●
	Formal Semantic Definition				●		○
	Concept Permanence	●		●	●	●	●
	Concept Nonredundancy	●	○	●	●	●	●
	Concept Nonambuiguity	●	○	●	●	●	●
Relationships	Monohierarchy (Tree)	●	●				
	Polyhierarchy	○	○	●	●	●	●
	DAG (Cycle-free)	●	●	○	●		

1.2.2 The Systematized Nomenclature of Medicine (SNOMED CT)

As shown in Figure 5-1, SNOMED CT is the most comprehensive set of clinical concepts. It is organized as a Directed Acyclic Graph (DAG) that builds on a model of well-formed concepts based on description logics. In addition to the partonomy and type relationships, it contains relationships that relate morphologies and anatomies with diseases. It is owned and approved by the College of American Pathologists and is available for free perpetual use in the USA through a license by the National Library of Medicine.

1.2.3 International Statistical Classification of Diseases (ICD-9, ICD9-CM, ICD-10)

ICD-9 and ICD-10 are detailed classifications of known diseases and injuries. ICD-10 is used world-wide for morbidity and mortality statistics, reimbursement systems and automated decision support in medicine. ICD-9 and ICD-10 are owned by the World Health Organization. The use of ICD-10 is subject to a licensing agreement with the WHO, though the terms are

generally free for research. ICD-9-CM is a clinical modification of the ICD-9 chiefly used for clinical billing in the USA.

1.2.4 Medical Subject Headings (MeSH)

MeSH is a terminology developed by the National Library of Medicine for the purpose of indexing journal articles and books in the life sciences [24]. It is used to index the MEDLINE/PubMed® article database. MeSH comprises about 23,000 descriptors and 150,000 supplementary concepts. MeSH is available electronically at no charge.

1.2.5 International Classification of Primary Care, Second Edition (ICPC-2)

ICPC-2 is a classification of about 1,000 terms of patient data and clinical activity in the domains of primary care. It has a biaxial structure consisting of (i) 17 clinical systems (chapters) and (ii) of 7 types of data (e.g. symptoms, diagnostic, screening and preventive procedures medication, treatment, test results, etc.).

1.2.6 Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV)

DSM-IV has been developed though a stringent experimental methodology to normalize the meanings of mental health disorder terms. It is published by the American Psychiatric Association. Its codes are defined to be compatible with ICD-9.

1.2.7 Logical Observation Identifiers Names and Codes (LOINC)

LOINC is a standard for identifying laboratory and clinical observations. It is approved by the American Clinical Laboratory Association and the College of American Pathologist. LOINC is not exactly an ontology. Rather, it supports the development of formal, distinct, and unique names corresponding to the description of the observation entities along six axes.

1.2.8 PaTO

To provide a unified framework for phenotypic representation, the Gene Ontology consortium has initiated the development of the Phenotype Attribute Ontology (PAtO) to reduce the structural barriers that limit the reuse of phenotypic databases. It consists of an ontology of phenotypic

attributes and an information model to communicate phenotypes across different communities as illustrated in Figure 5-2.

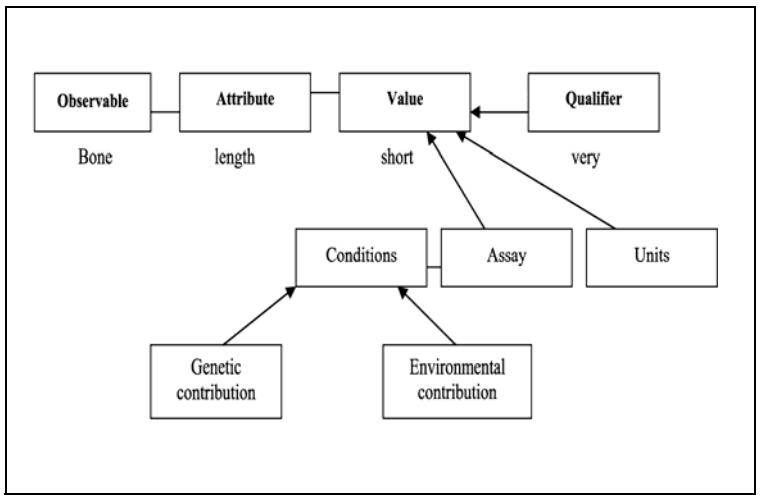

Figure 5-2. Simplified Phenotype Attribute Ontology Information Model

1.2.9 Unified Medical Language System® (UMLS®)

The UMLS of the National Library of Medicine is a semi-automated integration effort covering over one hundred terminologies [25-27]. It has been designed as a network (not a Directed Acyclic Graph) to honor relationships that it aggregates from source terminologies. The UMLS models the actual relationships among disparate concepts taken from information sources, achieving coordinated linkage of alternate encoding of data without the difficulty of pairwise integration. It also provides extensive semantic and lexical information about the terms associated with these concepts. It is one of the most comprehensive harmonized cross-mapping frameworks for biomedical terminologies currently available.

1.2.10 National Cancer Institute (NCI) Metathesaurus

The NCI Metathesaurus is another massive undertaking in the integration of terminologies. It has been developed by National Cancer Institute and contains 850,000 concepts mapped to 1,500,000 terms by over 4,500,000 relationships [28] and includes parts of the UMLS Metathesaurus.

2. INTEGRATION OF CLINICAL ONTOLOGIES

Phenotypes are poorly integrated across model organism database systems, the literature and human disease databases. Representation of phenotypic information is more complicated compared to biological data and consequently there are few data standards and data models for phenotypic information across species and within human repositories. In addition, the granularity of phenotypic data varies from database to database. Further, current methods for accessing phenotypic information across databases are inefficient.

The problem of integrating phenotypes across heterogeneous sources is compounded by a number of issues rooted in the complexity of phenotypic information and knowledge representation (ref. Section 5-1.1, Complexity of Representation of Clinical Phenotypes). These issues are due to differences in (i) definitions [2, 29, 30] and standards, (ii) compositionality and granularity [17-19, 31] (iii) biological scale [32], and (iv) context [14, 21, 33-35]. Moreover, the biomedical community has yet to reach a consensus on whether diseases, syndromes and behaviors are phenotypes, and the distinction between traits and phenotypes.

2.1 Integration of Ontologies' Concepts with the UMLS and Related Tools

The UMLS also has a number of related tools such as *MetaMap (MMTx)* for mapping terms to concepts in the UMLS Metathesaurus [36] and *Metamorphosys* for customizing the UMLS Metathesaurus (tailoring a subset of terminologies and their network of relationships by filtering the UMLS).

Mapping of various medical terminologies to the UMLS and other biomedical terminologies has been explored extensively [31, 37-48] and the utilization of semantics to interoperate terminologies was first proposed over a decade ago [49]. However, the attempted methods have had limited success. On average, they are only able to map 13 - 60% of the terms. These classes can be unified to create an integrated schema for the sources. Blake et al. have demonstrated that clustering techniques allow for the evaluation of candidate classes in different sources of terminologies. Hill et al. have manually integrated the Gene Ontology with external vocabularies [50]. While the use of description logics allow for automated evaluation of semantic relationships in the thesaurus, clustering techniques permit the evaluation of candidate classes in different sources that maybe unified in the global schema.

There are at least two problems associated with pre-coordination of terminologies for biomedical science: (i) *slow or rate-limiting updates* of the cross-index due to the resource intensive knowledge engineering, and 2) *computational ambiguity* of the reuse of a concept increases with the size of its terminology unless it is implemented with computable information about the context(s) of its usage. Further, part of the complexity lies in the variety of ways that a single biological concept may be represented [51]. As disparate systems often use the same information resources, it is imperative that redundancy be kept to a minimum in pre-coordinated systems. However, the issues of context and complexity make the pre-coordinated approach increasingly expensive and/or challenged for timeliness in the face of the escalating needs of biologists whose terminologies are undergoing accelerated updates. Additionally, different terminologies may represent the same concept in a very different way.

2.2 Integration via Information Models

There are few data model standards for combining phenotypic data across distinct databases. As shown in section 5-1, clinical phenotypes are usually specified in distinct sub-languages specific to scientific and professional subspecialties leading to restricted opportunities for relevant conceptual mappings across organisms or across disease databases. This is also compounded by the fact that clinical ontologies are generally developed independently of one another. Even when the sub-languages are similar and share the same structural representation, the *granularity* (detail) of their representation may still differ across databases. Indeed, ICD-9 comprises only about 25,000 clinical conditions while SNOMED CT describes over 100,000 clinical conditions. We briefly present two information models that may be used with clinical information: the broad HL7 and PAtO, specific to phenotypes.

2.2.1 HL7

Health Level Seven (HL7), is a volunteer-based and not-for-profit organization involved in the development of common data models for sharing clinical information. While version 2 of HL7 did not provide formalism for vocabulary support, version 3 now provides such structure.

2.2.2 PAtO Information Model

The PAtO information model presented in Figure 5-2 was originally intended to share phenotypic information across model organism databases and provides some insight on how to map clinical information with model organisms' phenotypes.

2.3 Integration of Clinical and Genomic Databases

Gene-Phenotype analyses are currently driven by quantitative trait loci studies requiring carefully curated pedigrees of patients of functional genomic studies. One of the limiting factors hindering the progress of clinical genomics discovery research is the lack of accurate and timely access to comprehensive gene-phenotypes networks associated with knowledge about biology and diseases due to the lack of integration across clinical and genomic databases. However, with the advent of the NIH Whole Genome Association studies, large volumes of well-organized clinical information are about to become available for high-throughput research.

Currently, while many genomic databases of model organisms contain some phenotypic information, phenotypes are often coded at different levels of granularity, in different formats, and with different aims [52]. Some efforts have been made in the integration and standardization of this data for sharing purposes. For example, the PhenomicDB [53] database provides a single portal for heterogeneous phenotypic information from a number of different model organisms and humans. It contains over 15,000 distinct phenotypic terms and 120,000 genotypes for the mouse and human species. Similarly, Gene2Disease was constructed over the Online Mendelian Inheritance in Men (OMIM) using text mining methods coupled with analysis of the chromosomal locations of diseases [54]. However, these systems make limited usage, if any, of clinical ontologies. In these two systems, the integration of phenotypes relies on the juxtaposition of the original lexical string of text in the same field across species. Thus a textual search for a concept may miss synonyms, as well as related or subsumed concepts. In contrast, the Mammalian Phenotype Ontology [55] is used by the Mouse Genome Database [56] to normalize representation across model organism databases (mouse and rat), via curation of annotations and a shared standard.

2.4 Integration with Natural Language Processing and Computational Terminologies

Among all natural language processing (NLP) technologies, MedLEE, developed by Friedman, has performed consistently and effectively in extracting clinical information, as evidenced by results of numerous independent evaluations [57-62]. BioMedLEE, is a NLP system derived from MedLEE and focused on parsing and coding gene-phenotype associations [63, 64]. In addition, lexico-semantic mapping of various medical terminologies to the UMLS and other biomedical terminologies has been explored extensively [31, 36, 38-44, 47, 49, 50, 65]. Previous NLP technologies would generally parse clinical data, but not encode them in clinical ontologies. New NLP systems for mining clinical narratives and coding in clinical ontologies are being developed. For example, the NIH National Center for Biomedical Computing "Informatics for Integrating Biology & the Bedside" (I2B2), headed by Isaac Kohane, is developing and distributing such a system as open source software [66].

3. DISCOVERY AND CLINICAL ONTOLOGIES

In the new millennium, the inception of the Gene Ontology (GO) precipitated a flurry of discovery methods and studies anchored on GO. Indeed, about one thousand scientific articles cite GO in their keywords. In comparison, about four thousand scientific articles cite ICD-9 and one thousand cite the UMLS or SNOMED. However, a dozen studies cite both GO and a clinical ontology, showing the tremendous opportunity for discoveries with ontology-anchored methods joining the biological and clinical scales.

3.1 Text Mining and Discovery

To overcome the limitations of manual annotation to create clinical phenotypic datasets, many informaticians have conducted high-throughput phenotype-genotype analyses by mining text on phenotype-genotype relationships from the scientific literature [67-75]. Recently, we have extended these approaches with semantic models of phenotypes to associate phenotypes with Gene Ontology Annotations in high-throughput [63], thus creating expressive and distinctive ternary relationships between genes, molecular classes and phenotypes.

3.2 UMLS and Discovery Systems

We and others have pioneered the integration of genomic databases with ontology-anchored clinical databases. Since clinical decision support systems like Quick Medical Reference (QMR) [76] contain densely coded descriptions of diseases, we hypothesized that they can be used as a proxy for clinical databases in genetic studies. To unveil systems biology properties of phenotypes via conducting genome-scale clustering analysis of phenotypes associated with diseases, we conducted two studies with QMR. In the first study, we applied terminological mapping and semantic techniques. Briefly, trait-disease-gene relationships buried in three databases (QMR, OMIM and SNOMED) were successfully integrated [77]. We also performed a clustering of OMIM's genes against QMR's traits of diseases and demonstrated a classification of diseases according to genes [77] comparable to the hierarchies found in ICD-9 or SNOMED. This study was followed up with the GenesTrace method, a large scale integrative study of ontology-anchored phenotypes from the UMLS and their statistical and semantic relationships to GO and model organism databases [78]. We were able to predict about three million phenotype-gene associations relationships between 22,040 phenotypic concepts in the UMLS and 16,894 gene products annotated using GO and its associated databases [78]. We validated our computed correlations by using OMIM's known gene-disease relationships as a gold standard. 30% of the predictions were found in OMIM, and similarly 9% of OMIM's relationships were found in GenesTrace [78]. Our methods provided direct links between genomic databases and clinically significant diseases through established clinical ontologies.

Recently, Butte and Kohane [79] conducted a study based on mapping results between phenotypically-related concepts in UMLS [80] and the microarray gene expression data from the NCBI's Gene Expression Omnibus (GEO) [81] using a term presence/absence method. Significantly expressed genes above a threshold were correlated with UMLS phenotypic concepts using a re-sampling-based multiple testing simulation generating 64,003 relations between 281 biomedical concepts and 7,466 genes. More importantly, their predictions were experimentally validated with microarray studies.

4. FUTURE CHALLENGES AND CONCLUSIONS

In this chapter, we highlighted the feasibility of computational approaches to conducting large-scale integrative studies anchored on clinical and biological ontologies and presented some realizations. Among various

strategies that could facilitate computational genomics studies, clinical ontologies are increasingly proving to be effective in integrating and organizing large amounts of phenotypic concepts. The success of the UMLS integration and reuse also attests the importance of ontologies in clinical research. Additionally, text mining techniques are increasingly relying on coded output in ontologies. The emerging field of high-throughput phenomics is likely to require the use of both clinical and biological ontologies as demonstrated in a few studies. Resources such as the UMLS, the NCI Metathesaurus, along with modern computational terminology tools will likely play an important role in the Semantic Web for Health Care and Life Sciences, encouraging the sharing and reuse of datasets. The Semantic Web offers a unique opportunity to commoditize access to these ontologies via OWL-based ontology servers and to provide tools automating the integration of databases coded in heterogeneous standards. Future interaction between the Semantic Web and clinical ontology is likely to proceed from the clinical concepts that have crisp definitions and require relatively simpler translational tables between distinct terminological standards, such as basic anatomical terms, simple lists of phenotypes, diseases and medications. Providing translation services via the Semantic Web is plausible in a near future.

DEFINITIONS

Terminologies: An ensemble of technical terms used in a specific domain.
Classification: A terminology with a systematic categorical arrangement.
Nomenclature: A comprehensive terminology enumerating extensively the terms used in a specific domain.
Ontologies: In this chapter, this term is used in its inclusive meaning in biology, which pertains to well-organized terminologies.

ACKNOWLEDGMENTS

We acknowledge the support of National Library of Medicine and of the following grants: the NIH/NLM 1K22 LM008308 (Semantic Approaches to Phenotypic Database Analysis), and the NIH/NCI 1U54CA121852-01A1 (National Center for the Multiscale Analysis of Genomic and Cellular Networks (MAGNet). We thank Lee Sam and Tara Borlawsky for their critical comments. This research was also supported in part by the Intramural

Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

REFERENCES

[1] Brunner H.G. and van Driel M.A. From syndrome families to functional genomics. Nat Rev Genet. 5(7): 545-51, 2004.

[2] Mahner M. and Kary M. What exactly are genomes, genotypes and phenotypes? And what about phenomes? Journal of Theoretical Biology. 186(1): 55-63, 1997.

[3] Musen M.A., Gennari J.H., Eriksson H., Tu S.W., and Puerta A.R. PROTEGE-II: computer support for development of intelligent systems from libraries of components. Medinfo. 8 Pt 1: 766-70, 1995.

[4] Rector A., Rossi A., Consorti M.F., and Zanstra P. Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN Organisation. Int J Med Inform. 48(1-3): 71-84, 1998.

[5] Campbell K.E., Das A.K., and Musen M.A. A logical foundation for representation of clinical data. J Am Med Inform Assoc. 1(3): 218-32, 1994.

[6] Friedman C., Huff S.M., Hersh W.R., Pattison-Gordon E., and Cimino J.J. The Canon Group's effort: working toward a merged model. J Am Med Inform Assoc. 2(1): 4-18, 1995.

[7] Bodenreider O. and Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform. 2006.

[8] Rubin D.L., Hewett M., Oliver D.E., Klein T.E., and Altman R.B. Automating data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML. Pac Symp Biocomput. 88-99, 2002.

[9] Embley D.W., Campbell D.M., Randy D.S., and Stephen W.L., *Ontology-based extraction and structuring of information from data-rich unstructured documents*, in *Proceedings of the seventh international conference on Information and knowledge management*. 1998, ACM Press: Bethesda, Maryland, United States.

[10] Honavar V., Silvescu, A., Reinoso-Castillo, J., Andoff, C., Dobbs, D. Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources. in Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources.2001

[11] Snoussi H., Magnin L., and Nie J.-Y. Heterogeneous web data extraction using ontologies. in Third International Bi-Conference Workshop on Agent-oriented information systems (AOIS-2001) Montréal, Canada,2001

[12] Yu H., Friedman C., Rhzetsky A., and Kra P. Representing genomic knowledge in the UMLS semantic network. Proc AMIA Symp. 181-5, 1999.

[13] Musen M.A. Dimensions of knowledge sharing and reuse. Comput Biomed Res. 25(5): 435-67, 1992.

[14] Rector A.L., Rogers J., Roberts A., and Wroe C. Scale and context: issues in ontologies to link health- and bio-informatics. Proc AMIA Symp. 642-6, 2002.

[15] Pole P.M. and Rector A.L. Mapping the GALEN CORE model to SNOMED-III: initial experiments. Proc AMIA Annu Fall Symp. 100-4, 1996.
[16] Elkin P.L., Tuttle M., Keck K., Campbell K., Atkin G., and Chute C.G. The role of compositionality in standardized problem list generation. Medinfo. 9 Pt 1: 660-4, 1998.
[17] Elkin P.L., Bailey K.R., and Chute C.G. A randomized controlled trial of automated term composition. Proc AMIA Symp. 765-9, 1998.
[18] Mays E., Weida R., Dionne R., Laker M., White B., Liang C., and Oles F.J. Scalable and expressive medical terminologies. Proc AMIA Annu Fall Symp. 259-63, 1996.
[19] Nelson S.J., Olson N.E., Fuller L., Tuttle M.S., Cole W.G., and Sherertz D.D. Identifying concepts in medical knowledge. Medinfo. 8 Pt 1: 33-6, 1995.
[20] Sujansky W. Heterogeneous database integration in biomedicine. J Biomed Inform. 34(4): 285-98, 2001.
[21] Oliver D.E., Rubin D.L., Stuart J.M., Hewett M., Klein T.E., and Altman R.B. Ontology development for a pharmacogenetics knowledge base. Pac Symp Biocomput. 65-76, 2002.
[22] Cimino J.J. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 37(4-5): 394-403, 1998.
[23] Cimino J.J. In defense of the Desiderata. J Biomed Inform. 39(3): 299-306, 2006.
[24] Nelson S.J., Johnston D., and Humphreys B.L., *Relationships in Medical Subject Headings*, in *Relationships in the organization of knowledge*, C.A. Bean and R. Green, Editors. 2001, Kluwer. p. 171-184.
[25] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32(Database issue): D267-70, 2004.
[26] Humphreys B.L., Lindberg D.A., Schoolman H.M., and Barnett G.O. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 5(1): 1-11, 1998.
[27] Lindberg D.A., Humphreys B.L., and McCray A.T. The Unified Medical Language System. Methods Inf Med. 32(4): 281-91, 1993.
[28] [cited; Available from: http://ncimeta.nci.nih.gov/indexMetaphrase.html.
[29] Strachan T. and Read A., *Human Molecular Genetics*. 2nd ed. 1999: Wiley-Liss. 574.
[30] Dawkins R., *The Extended Phenotype: The Long Reach Of The Gene*. 1982: Oxford University Press.
[31] Tuttle M.S., Suarez-Munist O.N., Olson N.E., Sherertz D.D., Sperzel W.D., Erlbaum M.S., Fuller L.F., Hole W.T., Nelson S.J., Cole W.G., et al. Merging terminologies. Medinfo. 8 Pt 1: 162-6, 1995.
[32] Blois M., *Information in Medicine: The Nature of Medical Descriptions*. 1984, Berkeley, California: University of California Press.
[33] Levy A., *Combining Artificial Intelligence and Databases for Data Integration*, in *Artificial Intelligence Today: Recent Trends and Developments*, M.a.V. Wooldridge, M, Editor. 1999, Springer: Berlin. p. 249-268.
[34] Friedman C., Hripcsak G., Shagina L., and Liu H.F. Representing information in patient reports using natural language processing and the extensible markup language. Journal of the American Medical Informatics Association. 6(1): 76-87, 1999.
[35] Krauthammer M., Johnson S.B., Hripcsak G., Campbell D.A., and Friedman C. Representing nested semantic information in a linear string of text using XML. Proc AMIA Symp. 405-9, 2002.

[36] Aronson A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 17-21, 2001.

[37] McCray A.T., Browne A.C., and Bodenreider O. The lexical properties of the gene ontology. Proc AMIA Symp. 504-8, 2002.

[38] Cimino J.J., Johnson S.B., Peng P., and Aguirre A. From ICD9-CM to MeSH using the UMLS: a how-to guide. Proc Annu Symp Comput Appl Med Care. 730-4, 1993.

[39] Tuttle M.S., Cole W.G., Sheretz D.D., and Nelson S.J. Navigating to knowledge. Methods Inf Med. 34(1-2): 214-31, 1995.

[40] Tuttle M.S., Sherertz D.D., Erlbaum M.S., Sperzel W.D., Fuller L.F., Olson N.E., Nelson S.J., Cimino J.J., and Chute C.G. Adding your terms and relationships to the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care. 219-23, 1991.

[41] Lussier Y.A., Shagina L., and Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proc AMIA Symp. 418-22, 2001.

[42] Masarie F.E., Jr., Miller R.A., Bouhaddou O., Giuse N.B., and Warner H.R. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 24(4): 379-400, 1991.

[43] McCray A.T., Srinivasan S., and Browne A.C. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 235-9, 1994.

[44] Rocha R.A., Rocha B.H., and Huff S.M. Automated translation between medical vocabularies using a frame-based interlingua. Proc Annu Symp Comput Appl Med Care. 690-4, 1993.

[45] Bodenreider O., Nelson S.J., Hole W.T., and Chang H.F. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp. 815-9, 1998.

[46] Fung K.W. and Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc. 266-70, 2005.

[47] Bodenreider O., Mitchell J.A., and McCray A.T. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. Proc AMIA Symp. 61-5, 2002.

[48] Lomax J. and McCray A.T. Mapping the Gene Ontology into the Unified Medical Language System. Comparative and Functional Genomics. 5: 354-361, 2004.

[49] Cimino J.J. and Barnett G.O. Automated translation between medical terminologies using semantic definitions. MD Comput. 7(2): 104-9, 1990.

[50] Hill D.P., Blake J.A., Richardson J.E., and Ringwald M. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. Genome Res. 12(12): 1982-91, 2002.

[51] Spackman K.A. and Campbell K.E. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. Proc AMIA Symp. 740-4, 1998.

[52] Biesecker L.G. Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. Clin Genet. 68(4): 320-6, 2005.

[53] Kahraman A., Avramov A., Nashev L.G., Popov D., Ternes R., Pohlenz H.D., and Weiss B. PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics. 21(3): 418-20, 2005.

[54] Perez-Iratxeta C., Wjst M., Bork P., and Andrade M.A. G2D: a tool for mining genes associated with disease. BMC Genet. 6: 45, 2005.

[55] Smith C.L., Goldsmith C.A., and Eppig J.T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol. 6(1): R7, 2005.

[56] Blake J.A., Eppig J.T., Bult C.J., Kadin J.A., and Richardson J.E. The Mouse Genome Database (MGD): updates and enhancements. Nucleic Acids Res. 34(Database issue): D562-7, 2006.

[57] Friedman C., Knirsch C., Shagina L., and Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp. 256-60, 1999.

[58] Hripcsak G., Friedman C., Alderson P.O., DuMouchel W., Johnson S.B., and Clayton P.D. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 122(9): 681-8, 1995.

[59] Hripcsak G., Kuperman G.J., and Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med. 37(1): 1-7, 1998.

[60] Jain N.L. and Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp. 829-33, 1997.

[61] Knirsch C.A., Jain N.L., Pablos-Mendez A., Friedman C., and Hripcsak G. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. Infect Control Hosp Epidemiol. 19(2): 94-100, 1998.

[62] Friedman C., Kra P., Yu H., Krauthammer M., and Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics. 17 Suppl 1: S74-82, 2001.

[63] Lussier Y.A., Borlawsky T., Rappaport D., and Friedman C. PhenoGO: a Multistrategy Language Processing System Assigning Phenotypic Context to Gene Ontology Annotations. Pacific Symposium on Biocomputing. 64-75, 2006.

[64] Friedman C., Borlawsky T., Shagina L., Xing H.R., and Lussier Y.A. Bio-ontology and text: bridging the modeling gap. Bioinformatics. 2006.

[65] Zeng Q. and Cimino J.J. Mapping medical vocabularies to the Unified Medical Language System. Proc AMIA Annu Fall Symp. 105-9, 1996.

[66] *2006 NCBC All Hands Meeting*. 2006: Bethesda, MD.

[67] Hafner C.D., Baclawski K., Futrelle R.P., Fridman N., and Sampath S. Creating a knowledge base of biological research papers. Proc Int Conf Intell Syst Mol Biol. 2: 147-55, 1994.

[68] Bajdik C.D., Kuo B., Rusaw S., Jones S., and Brooks-Wilson A. CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. BMC Bioinformatics. 6(1): 78, 2005.

[69] Yakushiji A., Tateisi Y., Miyao Y., and Tsujii J. Event extraction from biomedical papers using a full parser. Pac Symp Biocomput. 408-19, 2001.

[70] Perez-Iratxeta C., Bork P., and Andrade M.A. Association of genes to genetically inherited diseases using data mining. Nat Genet. 31(3): 316-9, 2002.

[71] Raychaudhuri S. and Altman R.B. A literature-based method for assessing the functional coherence of a gene group. Bioinformatics. 19(3): 396-401, 2003.

[72] Raychaudhuri S., Chang J.T., Sutphin P.D., and Altman R.B. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Res. 12(1): 203-14, 2002.

[73] Haft D.H., Selengut J.D., Brinkac L.M., Zafar N., and White O. Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics. 21(3): 293-306, 2005.

[74] Korbel J.O., Doerks T., Jensen L.J., Perez-Iratxeta C., Kaczanowski S., Hooper S.D., Andrade M.A., and Bork P. Systematic association of genes to phenotypes by genome and literature mining. PLoS Biol. 3(5): e134, 2005.

[75] Bodenreider O., *Lexical, terminological and ontological resources for biological text mining*, in *Text mining for biology and biomedicine*, S. Ananiadou and J. McNaught, Editors. 2006, Artech House. p. 43-66.

[76] Miller R.A. and Masarie F.E., Jr. Use of the Quick Medical Reference (QMR) program as a tool for medical education. Methods Inf Med. 28(4): 340-5, 1989.

[77] Lussier Y.A., Sarkar I.N., and Cantor M. An integrative model for in-silico clinical-genomics discovery science. Proc AMIA Symp. 469-73, 2002.

[78] Cantor M.N., Sarkar I.N., Bodenreider O., and Lussier Y.A. Genestrace: phenomic knowledge discovery via structured terminology. Pac Symp Biocomput. 103-14, 2005.

[79] Butte A.J. and Kohane I.S. Creation and implications of a phenome-genome network. Nat Biotechnol. 24(1): 55-62, 2006.

[80] National Library of Medicine. *Unified Medical Language System® Fact Sheet*. 2006 23 March 2006 [cited; Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html.

[81] Wheeler D.L., Church D.M., Edgar R., Federhen S., Helmberg W., Madden T.L., Pontius J.U., Schuler G.D., Schriml L.M., Sequeira E., et al. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res. 32(Database issue): D35-40, 2004.