Health Informatics

Peter L. Elkin *Editor*

Terminology and Terminological Systems

EXTRA MATERIALS
extras.springer.com

Springer

Health Informatics

Peter L. Elkin
Editor

Terminology and Terminological Systems

Springer

Editor
Peter L. Elkin, M.D., MACP, FACMI
Physician, Researcher and Author
New York, NY
USA

Additional material to this book can be downloaded from http://extras.springer.com

ISBN 978-1-4471-2815-1 ISBN 978-1-4471-2816-8 (eBook)
DOI 10.1007/978-1-4471-2816-8
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2012937652

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

It is a sunny day. It was not always sunny in the life of the dedicated physician from southern Pennsylvania. The goals of a lifetime exist mostly in the mind of those who hold them dear. Lifetimes of adjustments and sacrifices culminate in a sense of purpose, honor, and dedication. Goals like positively influencing the lives of others are relative in time and space to one's own vision. Happiness predicated on a consistent view of the present and future is irreconcilable with the ever more rapid evolution of the present. Time once the close friend and trusted ally of the thoughtful becomes a consistent reminder of the limits of Man. Dreams may be infinite, but our capacity to implement them is not. Trusting in the prioritization of others oft leads to a constant pressure toward mediocrity. Comfort can be found in the familiar, so does mediocrity yield the same warm complacency. Struggle though empowering does not lead to happiness. Accomplishments being relative by almost all standards, our struggles are in vane.

In this spirit, I submit this work whose timeliness will be limited and whose goal will need to evolve with the sophistication of the student. As a teaching text, it melds history with theory and logic. As a vision, it opens the mind to a greater potential, and it is my profound hope that in the telling I can stimulate thought and progress in the representation of language and the instantiation of thought.

I am continually inspired and enthralled with the human mind's ability to imagine a better future, and I am of the belief that we can create any manifest future that we can imagine. My hope is that this textbook inspires you to make a difference in people's lives.

Herein, we imagine improved healthcare through better organization, dissemination and reasoning about real patient problems, routinely elevated by ubiquitously available medical and ontological knowledge.

I dedicate this book to the two great women in my life, my mother Lorretta Elkin who spent her life in dedication to truth and honesty and her children, along with my wife Margaret Ann who has dedicatedly inspired this admitted workaholic.

Peter L. Elkin

Foreword

This book addresses a deep and pervasive paradox. On the one hand, "terminology" is central to healthcare. Simply put, education, patient records, laboratory testing, procedures, medication, reimbursement, assessment of quality, regulatory compliance, and research could not exist without it. On the other hand, "awareness" of terminology as a rate-limiting resource is low, generally. That terminology is something that could be missing or present, or good or bad, or the subject – in current parlance – of "best practices," just has not occurred to many outside the relatively narrow confines of medical informatics. And those within that field have only the literature – spread across many journals and sources – to guide them.

The pages that follow put a stake in the ground. As is explained, terminology comes from somewhere – both historically (millennia) and at present (from various authorities), and it is going somewhere (at this writing, "Meaningful Use" – in the United States – will require it, and attempts to create international terminology are making progress). Further, terminology is already intimately connected with technology – information technology in particular. As implied in many sections of this book, terminology empowers computers as healthcare strives to catch up with the productive use of information technology in other fields.

One reason for the apparent shortfall in the use of information technology in healthcare and in biomedicine – again relative to other domains – is the healthcare and biomedical terminology challenge. First, the magnitude requirements for just lab tests, medications, procedures, and diagnoses dwarf that found in any other significant context, and the burgeoning "naming" requirements for the study and use of "genotype-to-phenotype" links have required the development of distributed provenance mechanisms. Second, the terminologies discussed in this book evolve continuously; authoritative repertoires of names for lab tests and medications change every day, and the names of procedures and diagnoses change annually, and sometimes more often. Third, seemingly obvious terminology-based queries prove to be research problems currently: examples are "What is our (local) experience with patients like mine?" "Which enterprise evidences 'best practice' for diagnosis X?" and "What can we learn from aggregating data from multiple sites that we cannot learn from the individual sites?"

Those attempting to address these challenges – locally, nationally, or internationally – need to appreciate the ideas discussed in this book. For example, those helping to manage care or research enterprises will be called upon to

supervise the management of terminology as an asset. Narrowly, terminology will be an asset because it will be the only way to "normalize" enterprise care and research data; more broadly, terminology may make explicit what makes the enterprise unique. At this writing, an emerging enterprise imperative is to integrate the terminology used to care for patients with the terminology used to conduct research relevant to those patients; standard terminology will need to coexist with novel names for research driven observations. On an individual and enterprise level, local terminology innovation will need to be coordinated with evolving extraproject, and extra-enterprise, terminology authorities.

This book could not have been written 20 years ago. Twenty-five years ago, the notion that terminology should be concept-based was all but unknown in healthcare; now, almost all important terminologies are at least partly concept-based. In parallel, because there was no general model of what a terminology was or should be, there were no tools to support terminology development and maintenance. Steady progress since then has improved both terminology content and the technology and processes used to sustain that content. This is the first book devoted to that story.

Students, practitioners, or managers who absorb the material here will have an advantage over their peers who lack it. Near term, terminology will become the bottleneck for the deployment of innovation and the assessment of quality in healthcare; opening that bottleneck will require understanding of elements of this text. Midterm, terminology will be an asset to be leveraged in care and research; for example, interenterprise clinical and research data aggregation – a central topic here – will become a dominant paradigm. Longer-term, but within the professional lives of informatics students who will learn from this book, terminology development and maintenance will become a distributed activity undertaken by individuals and by "crowds" (as in crowd-sourcing). As with software, maintenance will come to dominate creation as an intellectual and operational activity. Emerging "best practices" in terminology – based on many of the ideas covered in this book – will specify how local, distributed, national, and international maintenance should be undertaken productively.

Ridgefield, CT, USA — Mark Samuel Tuttle
Nashville, TN, USA — Steven H. Brown

Series Preface

This series is directed to healthcare professionals leading the transformation of healthcare by using information and knowledge. For over 20 years, Health Informatics has offered a broad range of titles: some address specific professions such as nursing, medicine, and health administration; others cover special areas of practice such as trauma and radiology; still other books in the series focus on interdisciplinary issues, such as the computer based patient record, electronic health records, and networked healthcare systems. Editors and authors, eminent experts in their fields, offer their accounts of innovations in health informatics. Increasingly, these accounts go beyond hardware and software to address the role of information in influencing the transformation of healthcare delivery systems around the world. The series also increasingly focuses on the users of the information and systems: the organizational, behavioral, and societal changes that accompany the diffusion of information technology in health services environments.

Developments in healthcare delivery are constant; in recent years, bioinformatics has emerged as a new field in health informatics to support emerging and ongoing developments in molecular biology. At the same time, further evolution of the field of health informatics is reflected in the introduction of concepts at the macro or health systems delivery level with major national initiatives related to electronic health records (EHR), data standards, and public health informatics.

These changes will continue to shape health services in the twenty-first century. By making full and creative use of the technology to tame data and to transform information, Health Informatics will foster the development and use of new knowledge in healthcare.

About the Authors

Dr. Steven H. Brown, M.D., MS, FACMI earned an A.B. Degree in Biology from Brown University in 1981, an M.D. from Brown University in 1987, and an M.S. in Biomedical Engineering from Vanderbilt University in 1998. His post-doctoral training in Internal Medicine was performed within the Emory University system. Dr. Brown completed an NIH/NLM fellowship in Applied Medical Informatics between 1994 and 1996.

Dr. Brown has been with VA since 1996. He has served as the Chief Information Officer at the Nashville VA Medical Center and the Tennessee Valley Healthcare System. He has worked for the Office of Information's Health Information Architecture group in the area of data standardization and knowledge representation. He was the founding Director of the Compensation and Pension Exam Program between 2001 and 2008. Dr. Brown is presently the Director of the Knowledge Based Systems (KBS) Office, a new program office within Veterans Health Administration Office of Informatics and Analytics. The KBS was created to help extend past VA informatics successes by infusing clinical informatics expertise into VHA healthcare decision making, strategic planning, and delivery. He is an Associate Professor of Biomedical Informatics at Vanderbilt University and Director of the Tennessee Valley Healthcare System special fellowship in medical informatics.

Dr. Brown's career in informatics spans 30 years and three academic institutions. He has authored or coauthored over 70 peer-reviewed medical informatics manuscripts and is one of only five VA informaticians elected to the American College of Medical Informatics. Among his achievements are the Vice President's "Hammer" award and the Secretary's "Scissors" award for the Compensation and Pension Record Interchange (CAPRI) software project.

Dr. Brown maintains a VA primary care practice and is Board Certified in Internal Medicine. He routinely uses VA clinical computing systems and understands their strengths and opportunities for improvement.

Dr. Peter L. Elkin, M.D., MACP, FACMI has served as a tenured Professor of Medicine at the Mount Sinai School of Medicine. In this capacity he was the Center Director of Biomedical Informatics, Vice-Chairman of the Department of Internal Medicine, and the Vice-President of Mount Sinai hospital for Biomedical and Translational Informatics. Dr. Elkin has published over 120 peer-reviewed publications. He received his Bachelor of Science from Union College and his M.D. from New York Medical College. He did his Internal Medicine residency at the Lahey Clinic and his NIH/NLM sponsored fellowship in Medical Informatics at Harvard Medical School and the

Massachusetts General Hospital. Dr. Elkin has been working in Biomedical Informatics since 1981 and has been actively researching health data representation since 1987. He is the primary author of the American National Standards Institute's (ANSI) national standard on Quality Indicators for Controlled Health Vocabularies ASTM E2087, which has also been approved by ISO TC 215 as a Technical Specification (TS17117). He has chaired Health and Human Service's HITSP Technical Committee on Population Health. He served as the co-chair of the AHIC Transition Planning Group. Dr. Elkin is a Master of the American College of Physicians and a Fellow of the American College of Medical Informatics. He chairs the International Medical Informatics Associations Working Group on Human Factors Engineering for Health Informatics. He was awarded the Mayo Department of Medicine's Laureate Award for 2005. Dr. Elkin is the index recipient of the Homer R. Warner award for outstanding contribution to the field of Medical Informatics.

Elizabeth Lumakovska, **MPA** is currently employed at the American Medical Association (AMA) as a Mapping and Terminology Consultant for the CPT Medical Informatics and Healthcare Strategy department. She has worked for the AMA since 1999. Prior positions she has held at the AMA include: Director, CPT Editorial Research and Development; Senior Terminology Analyst, CPT Electronic Content Development; and Coding Consultant, CPT Research and Development, Health Information Specialist, CPT Education and Information Services. Her adjunct employment includes teaching classes at Indiana University Northwest. She has been working in the medical field since 1991 and her past employment experiences, aside from the AMA, include: Quality Improvement Analyst, St. Anthony Medical Center; Health Information Manager, Children Memorial Health Care Resources; Human Resources Coordinator, BHM Health Associates Home Care; and Office Manager, Djurovic Medical Clinic.

Elizabeth's educational background consists of earning her Master Degree, Public Affairs – Health Administration Concentration; Bachelor of Science Degree, Health Services Management; Associate of Science Degree, Health Information Technology; and Graduate Certificate, Public Management from Indiana University. Additionally, she attained the Registered Health Information Technician (RHIT) credentials from the American Health Information Management Association (AHIMA) and the Certified Professional in Electronic Health Records (CPEHR) certification from the Certification Commission for Healthcare Information Technology (CCHIT).

Marjorie Rallins, **DPM**, **MA** is the Director of Measure Specifications, Standards and Informatics for the AMA and the AMA-convened PCPI. She leads the effort to ensure PCPI quality measures and supporting specifications integrated with EHRs and other HIT. Dr. Rallins and the highly skilled team has developed the EHR specifications for many of the quality measures in the CMS EHR Incentive Program (Meaningful Use) Stage I and are currently involved in preparing quality measure specifications for Meaningful Use Stages II and III.

Dr. Rallins has an extensive background in health care informatics and vocabulary standards. She is involved in a number of health information technology efforts including the vocabulary work group of the Health Information

Technology Standards Committee, the content committee of the International Health Technology Standards Development Organization (IHTSDO), and the National Quality Forum's Terminology Expert Panel.

Prior to coming to the AMA, Dr. Rallins was the Director of Clinical Editors for the College of American Pathologists where she led the terminology operations effort for SNOMED International. She received a Doctor of Podiatric Medicine degree from the Illinois College of Podiatric Medicine.

Dr. S. Trent Rosenbloom, M.D., MPH, FACMI is an Associate Professor of Biomedical Informatics with secondary appointments in Medicine, Pediatrics, and the School of Nursing at Vanderbilt University. He is a board certified Internist and Pediatrician who earned his M.D., completed a residency in Internal Medicine and Pediatrics, a fellowship in Biomedical Informatics, and earned an M.P.H. all at Vanderbilt. Since joining the faculty in 2002, Dr. Rosenbloom has become a nationally recognized investigator in the field of health information technology evaluation. His work has focused on studying how healthcare providers interact with health information technologies when documenting patient care and when making clinical decisions. His work has resulted in lead and collaborating authorship on 50 peer-reviewed manuscripts, which have been published in the *Journal of the American Medical Informatics Association*, *Pediatrics*, *Annals of Internal Medicine*, and *Academic Medicine*, among others. In addition, he has authored and coauthored 5 book chapters and numerous posters, white papers, and invited papers. He has been a committed member of the principal professional organization in his field, the American Medical Informatics Association (AMIA). As a result of his research success and service to AMIA, he was the annual recipient of the competitive AMIA New Investigator Award in 2009. In addition, Dr. Rosenbloom has participated in study sections for the National Library of Medicine and the Agency for Healthcare Research and Quality's Healthcare. He has also participated as a member of the HL7 Pediatric Data Special Interest Group and the American Academy of Pediatrics' Council on Clinical Information Technology. In addition, he is an active reviewer for several journals covering general medicine, pediatrics, and biomedical informatics.

Jennifer Trajkovski, MJ, RHIT, CHC is a Senior Policy Analyst for the Specifications, Standards, and Informatics department within the Performance Improvement Group at the American Medical Association (AMA). Trajkovski works on the development and maintenance of technical specifications for performance measures developed by the AMA convened Physician Consortium for Performance Improvement® (PCPI). She has extensive terminology experience, in particular Current Procedural Terminology® (CPT). She manages the terminology maintenance process for PCPI measure specifications. Prior to joining the AMA, she served as a senior consultant for a health care consulting firm where she specialized in regulatory compliance. She also worked previously for the AMA as a coding specialist in the CPT Product Development department. Trajkovski earned a Master of Jurisprudence and Health Policy degree from Loyola University Chicago.

Mark Samuel Tuttle, AB, BE, FACMI is on the Board of Directors of Apelon, Inc. where he was a co-founder. Before helping to start Lexical Technology, which became Apelon, he taught computer science at the University

of California Berkeley (UCB) and medical information science at the University of California San Francisco (UCSF). With Scott Blois, M.D., Ph.D., Mark led the UCSF team that won one of four initial awards from the United States National Library of Medicine to help build the Unified Medical Language System (UMLS). As part of this project Mark proposed and led the extramural development of what became the UMLS Metathesaurus. He led or worked on related projects for the US National Cancer Institute, US Veterans Affairs, US Department of Defense, US Health and Human Services, Kaiser Permanente Healthcare, and several private companies. As Vice President of Strategy at Apelon, he presented and published his work nationally and internationally, serving for many years on the Program Committee of the American Medical Informatics Association. He was elected as a Fellow of the American College of Medical Informatics (FACMI) in 1993. Mark received a liberal arts degree from Dartmouth College, and a professional engineering degree from the Thayer School. He also studied information science in the Masters program at Thayer and computer science in the Ph.D. program at Harvard. At Harvard he was the Head Teaching Fellow for an introductory computer class that became the third largest class at the University. At UCB Mark started a class titled "The Art and Science of Computing" that enrolled more than 800 students per year. More recently, Mark was a Teaching Assistant for the world's largest online class Stanford's Introduction to Artificial Intelligence, and he contributed to PL Elkin's informatics textbook Terminology and Terminology Systems.

Acknowledgements

Dr. Elkin would like to acknowledge his coauthors of the textbook and his many teachers over the course of his educational development.

Contents

Contributors

Steven H. Brown, M.D., MS, FACMI Department of Veterans Affairs and Vanderbilt, Department of Biomedical Informatics, Nashville, TN, USA

Peter L. Elkin, M.D., MACP, FACMI Physician, Researcher and Author, New York, NY, USA

Elizabeth Lumakovska, MPA CPT Medical Informatics and Healthcare Strategy, American Medical Association, Chicago, IL, USA

Marjorie Rallins, DPM, MA Performance Improvement, American Medical Association, Chicago, IL, USA

S. Trent Rosenbloom, M.D., MPH, FACMI Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

Jennifer Trajkovski, MJ, RHIT, CHC Performance Improvement, American Medical Association, Chicago, IL, USA

Mark Samuel Tuttle, AB, BE, FACMI Apelon, Ridgefield, CT, USA

Introduction

1

Peter L. Elkin and Mark Samuel Tuttle

This is a textbook designed to help students learn about healthcare terminology and other forms of controlled representations. Students will learn the purpose, organization, and use of these representations and as a result should be able to leverage them in clinical settings. Toward this end, each chapter will be use-case driven. In turn, the use cases will generate the examples. Each chapter will be followed by a set of questions for use as part of an informatics course or in support of self-study. Half of the answers will appear in the book; the rest will appear in an accompanying teachers' guide.

Terminologies in healthcare gained popularity as a method for representing knowledge about clinical events and healthcare data. Because of this, many terminologies came to be used for reimbursement and regulatory compliance. It is important to remember the order in which this happened. More generally, language used in healthcare evolved to be as expressive as clinicians thought necessary to describe the clinical events and patients that they were seeing. Evolution in medical descriptions drives the evolution of terminologies. Terminologies are also used by statistical organizations and authorities to count clinical events (e.g., in mortality registries). For the purposes of this book, "terminology" and "vocabulary" will be used interchangeably and be considered synonyms. Some writers chose to distinguish the two terms, but this writer believes that these distinctions serve more as a barrier to understanding than as a help to the informatics student. A concept is defined as the embodiment of some specific meaning and not a code or character string. A term is defined as a word or words corresponding to one or more concepts.

The goal of healthcare terminologies was and is to aggregate patient descriptions by meaning. The desired aggregations require that the terminologies be unambiguous and nonredundant. Unambiguous means that the concepts named in a given terminology each have a unique meaning. The abbreviation MS can stand for mitral stenosis in one context and multiple sclerosis in another. This is an example of an ambiguous term. A concept being nonredundant means that there are no two concepts in a given terminology have the same meaning. If we created a terminology which included two concepts one for "heart attack" and the other for "myocardial infarction," each with their own concept identifier (which we will speak more about in later chapters), we would have created redundant concepts in our terminology. In other words, the two quoted strings name the same concept and therefore should be synonyms.

Most terminologies in healthcare began as lists of categories or as a classification (e.g., The London Bills of Mortality). These were developed from lists of labels for concepts (ideas)

P.L. Elkin, M.D., MACP, FACMI (✉)
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

M.S. Tuttle, AB, BE, FACMI
Apelon, Ridgefield, CT, USA

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_1, © Springer-Verlag London 2012

expressed as noun phrases. When plain language definitions (e.g., English language) were assigned to the concepts, the meaning of the concept was fixed so that more than one reader would obtain the same meaning when reading the concept. For example, "myocardium" could be defined as "the muscle of the heart."

When compound expressions were developed where parts of the expression were defined, we call these systematic definitions. For example, "myocardial infarction" is "death due to lack of blood flow of the muscle of the heart." Then, it follows that "infarction" must be "death of some tissue due to a lack of blood flow to that tissue or organ."

It is meaning rather than words that is the important binding concept. So if "jaundice" were replaced by "icterus" (i.e., the Latin representation being replaced by the word derived from Greek), the systematic nature of the definition would still hold. This brings us to one of the central reasons for controlled vocabularies. Synonyms are associated with a concept and are linked together in the best case by meaningless identifiers. This will be discussed in greater detail later in the book. Again, this book is about units of meaning, their names, and the ways they can be created, maintained, and used in healthcare. It is not about words and their meanings.

The other great benefit of controlled representations is the ability of subject matter experts to organize the concepts within the terminology into subtypes. These subtypes can form hierarchies. For example, all "healthcare concepts" may have a subtype of "disorders" which may have a subtype of "cardiovascular disorders" which may have a subtype of "myocardial infarction." These hierarchies can help to retrieve all instances of a concept in a database or any instance of one of its subtypes or as we often refer to them as children (i.e., descendants).

This brings me to a discussion of knowledge. The study of knowledge is epistemology. The aspect of epistemology that will be discussed in this textbook is the representation of this knowledge. Knowledge representation is important for all aspects of research and for our understanding and continuous quality improvement of clinical care.

There are only three types of knowledge in the world. First, there is ontologic knowledge which is definitional (i.e., concepts from a terminology and their formal and systematic definitions) and of which we will focus considerable energy in this textbook. Second there is assertional knowledge, which are facts or axioms. An example of such a fact is that patients who develop a flail mitral leaflet (one that is no longer attached by its cordae to the heart) can develop flash pulmonary edema, or renal dialysis can be used to treat patients with end-stage renal disease. The last type of knowledge is instance data. If I developed a case of pneumonia and my physician recorded the information in my electronic health record, they would be recording that I had an instance of the concept of pneumonia. My pneumonia may not be the same as someone else's pneumonia as it may be caused by a specific bacteria and be located in a specific region of my lung(s). All knowledge can be categorized as either ontological knowledge, assertional knowledge, or instance knowledge.

In this book, we will discuss the history of terminologies, the mathematical and ontological basis of terminologies, terminological theory, specific terminological systems which are used to develop, maintain, or utilize terminologies, and terminologies as applied to domains, and we will discuss specific terminologies used in healthcare.

This book is intended to highlight as examples specific terminologies and terminological systems. It is not intended to be an exhaustive account of all terminologies and terminological systems. Nor does it account for them in order of importance. There are many important terminological efforts that will not be discussed in this text. However, we do try to cover enough terminologies to highlight the major issues of import in the development, maintenance, and use of controlled terminologies in healthcare practice, education, and research.

As we stated earlier, this book is use-case driven. The use case that will be used in the book is:

The sun was shining and Mr. John and Mrs. Jennifer Workalot are taking their first family vacation in three years. They are accompanied

by their two children: Michael, a very active 8 year old, and Rachael, a very bored four and a half year old. They are driving to Disney World from their home in Nashville, Tennessee. On the second day of their trip while passing through Georgia, the weather worsens and heavy rains appear. Not wanting to miss Mickey and not willing to tolerate further driving with two bored and active children, the Workalots continue driving into the night. A truck veers into their lane and they swerve off the road hitting a pole. They are rushed to Grady Memorial Hospital where Mr. Workalot is found to have an epidural hematoma requiring evacuation. Before surgery, his Vanderbilt records are obtained and the records are sent in an interoperable form such that Grady's expert system can run off of the data from the Vanderbilt record. This alerts the Grady neurosurgeon that John is allergic to penicillins and cephalosporins which otherwise might have been given during the case. John also has a family history of malignant hyperthermia, so care must be taken in choosing his anesthetic agents.

Jennifer has diabetes mellitus with a history of diabetic nephropathy, and even though she complains of no discomfort, her Grady clinicians check her legs and note that she has acute swelling of the left calf with a bluish discoloration of her ipsilateral toes. This is found by orthopedics to have a pressure of 60 mmHg and as such constitutes a compartment syndrome, and she is taken to the OR for a fasciotomy, thereby saving her leg from dangerous ischemia. Knowing her diabetic history will assist the clinicians in their management of her diabetes perioperatively.

Michael is inconsolable. From his pediatric records from Vanderbilt, we note that he is considered healthy, but due to his behavior and separation from his parents, he is visited by child psychiatry. It is determined that Michael is suffering from an acute stress reaction, and they wish to prescribe a selective serotonin reuptake inhibitor, namely, escitalopram (Lexapro). However, Michael's Vanderbilt records show his DNA sequence data with a polymorphism associated with nonfunctioning CYP3A4 enzyme indicating that Michael would likely be a poor metabolizer of escitalopram. Therefore, he is placed on sertraline (Zoloft) which is metabolized by the p450 CYP2D6 and CYP2C19 pathways, but is not metabolized via the CYP3A4 enzyme. This medication decision avoids subjecting Michael to potentially serious medication side effects that may have occurred through inadvertent overdosing of escitalopram.

Rachael presents with a painful right lower extremity having hit her knee on the seat in front of her. The leg is swollen from the knee to her toes and ultrasound reveals a deep venous thrombosis. Rachael's Vanderbilt record shows that she has a history of chronic bronchitis and she has been plagued by frequent infections. She is on chronic suppression with erythromycin. The clinician institutes heparin therapy, and as the clinician starts to order warfarin a warning of a potentially severe drug–drug interaction between erythromycin and warfarin is displayed. This leads to a change in management potentially avoiding a major bleeding episode.

After a week of well directed care, the Workalots are all released from Grady Memorial in good condition. This year's vacation did not work out well, but the next year the family traveled by air to Orlando and had a wonderful and healthy vacation.

Questions

1. Which of the following is not a type of knowledge?
   (a) Instance data
   (b) Ontological
   (c) Representational
   (d) Assertional
2. True or false, terminologies are used to populate mortality registries?
3. What is the distinction between a controlled vocabulary and a terminology?
   (a) A controlled vocabulary does not have concepts organized hierarchically.
   (b) A terminology does not have concepts organized hierarchically.
   (c) Controlled vocabularies are controlled by limiting what topics they can represent.
   (d) They are synonymous.

4. What is the goal of creating a controlled vocabulary?
   (a) To keep your field from being easily understood by others
   (b) To aggregate information by meaning
   (c) To contain the assertional knowledge known by experts in the field
   (d) To keep track of specific patient's problems
5. A concept is unambiguous if:
   (a) It is able to be understood in two distinct ways.
   (b) It has more than one meaning.
   (c) Multiple readers can reasonably interpret the meaning of the concept differently.
   (d) It has only one meaning.
6. Which of the following pairs of concepts are redundant?
   (a) Heart/heart attack
   (b) Myocardial infarction/acute myocardial infarction
   (c) Heart muscle/myocardium
   (d) Myocardial infarction/cardiovascular disease
7. An example of a subtype hierarchy of concepts is:
   (a) Entity/disease/cardiovascular disease/acute myocardial infarction/myocardial infarction
   (b) Acute myocardial infarction/myocardial infarction/cardiovascular disease/disease/entity
   (c) Disease/cardiovascular disease/myocardial infarction/acute myocardial infarction/entity
   (d) Entity/disease/cardiovascular disease/myocardial infarction/acute myocardial infarction
8. Systematic definitions:
   (a) Are created by using terminological systems
   (b) Are written by a computer
   (c) Follow a compositional method of defining the concepts
   (d) Follow rules for defining the identifiers associated with concepts
9. True or false, synonyms associated with concepts are given separate concept identifiers?
10. Which of the following best describes the benefits of controlled terminologies?
    (a) The aggregate information by meaning.
    (b) They contain synonymy so that they can recognize multiple terms that represent the same concept or idea.
    (c) They are arranged in hierarchies which allows users to get information stored that is related to a concept and all its children.
    (d) All of the above.

History of Terminology and Terminological Logics

2

Peter L. Elkin and Mark Samuel Tuttle

Introduction

Although many see William Farr from the nineteenth century as the father of terminology and classification, we can find evidence in the work of Hippocrates that earlier efforts were underway and meaningful.

Hippocrates was born in 460 BCE to Heraclides, a physician. In his contributions to the Corpus of Hippocrates, he organized medical knowledge into categories such as cautery or excision. He wrote disease-oriented treatise based on organ systems such as lung cancer and lung empyemas. He organized treatments by disorder. This type of systematic organization of health concepts can be said to be the beginning of controlled healthcare vocabularies.

Aristotle, approximately 100 years later, credited Hippocrates with the first organized thinking in health care. Aristotle himself is credited with the development of the first formal logic [1]. This began with categorization which concerned itself with the natural naming of things and extended itself in his volume names On Interpretation which defined the language and form of propositional statements and their elementary relations. He also wrote volumes on Prior Analytics, Posterior Analytics, Topics, and On Sophistical Refutations which discussed dialects. The first three volumes gave the grammar and rules for the construction of the language of logic. This was also called symbolic logic. These volumes were organized into a treatise named the Organon.

Aristotle was a student of Plato in the school of Athens. Plato looked for Universals and applied principally deductive reasoning to reach his conclusions. Aristotle extended this work to define not only Universals but also Particulars. This included not only deductive but also inductive reasoning. Here, we see the first use of logics to define Instance knowledge. This also implies the ability to direct the development of classification using real-world data rather than universal forms. Universals can be things such as an Apple or can be a property such as the shape of the Apple. Here there is a general shape of an apple and the shape of a specific apple (an instance of the universal apple). Here Aristotle emphasizes the need to represent knowledge about apples with different shapes. He also discussed relations. If an apple falls from the tree, it may have some relationship in space to the roots of the tree.

Aristotle defined the term "natural philosophy" to define the development of classification using phenomenon observed from the natural world. This has been extended to cover biology, health, and health care. He dissected many animals including fertilized eggs through maturation and systematically described what he observed.

Aristotle defined what he called "term logic" which later became known as propositional logic.

P.L. Elkin, M.D., MACP, FACMI (✉)
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

M.S. Tuttle, AB, BE, FACMI
Apelon, Ridgefield, CT, USA

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_2,

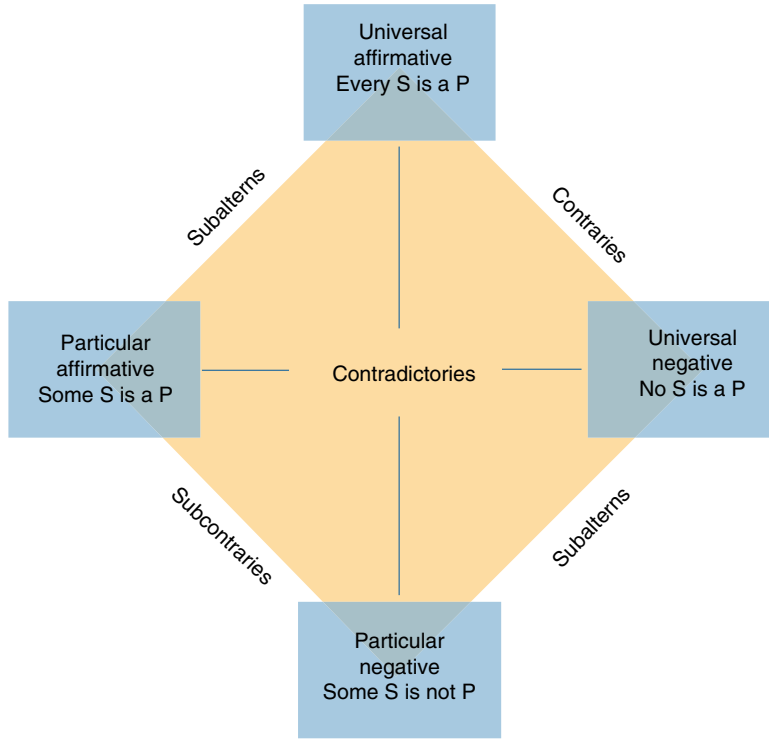


Fig. 2.1 Aristotle's square of opposition

Here, he defined the term as an entity or something. A proposition is defined as consisting of two terms where one is either affirmed or denied. The syllogism is where one proposition (the conclusion) follows from two other propositions (the premises) [2]. Propositions come in different types:

A-Type: Are Universal and Affirmative such as "All bleeding stops."

I-Type: Are Particular and Affirmative such as "Some people choose careers in Informatics."

E-Type: Are Universal and Negative such as "No one is immortal."

O-Type: Are Particular and Negative such as "Some students will not have trouble with this material."

These were called the square of opposition (see Fig. 2.1).

The term has evolved in modern thinking to the concept. Here, the notion is that the concept is an abstract representation of the thing or abstract notion such as good or evil. The concept should be language independent, have meaningless identifiers, and can be formally defined.

William Farr, a British Epidemiologist, is often regarded as the father of medical statistics [3]. After his wife died of tuberculosis in 1836, he took a job as the first compiler of scientific abstracts. His department was responsible for cataloging and recording the causes of death categorized by occupation. He called this catalog *Vital Statistics* and was elected as president of the Royal Statistical Society. This eventually became the London Bills of Mortality which was the precursor of the International Classification of Diseases (ICD). We use ICD9 Clinically Modified or ICD9-CM for morbidity coding and ICD10 for mortality coding in the USA today. ICD11 is currently under construction by the World Health Organization (WHO).

Syntactic, Semantic, and Pragmatic Interoperability. We tested the scale by having five medical Informaticians rate a set of ANSI standard specifications, and we report the interrater variability of the interoperability rating scheme. We learned that some elements of the scale presented more difficulty for our reviewers, and based on our findings we present a final version of the

interoperability scale in our discussion. Our interoperability rating ontology has high inter-rater reliability and is a relatively simple mechanism for comparing the levels of interoperability afforded by different specifications or the same specification over multiple versions.

Until the 1920s, logic and mathematics was often considered spiritual not scientific. Since the time of Pythagoras, mathematics was considered a revelation of the divine order. In Principia Mathematica (Russell and Whitehead), the authors demonstrated that mathematics was logical. Logical positivism was then applied to science and psychology.

Adolphe Quetelet was one of the most influential social statisticians of the nineteenth century. His applications of statistical reasoning to social phenomena profoundly influenced the course of European social science. Quetelet had come to be known as the champion of a new science, dedicated to mapping the normal physical and moral characteristics. Quetelet called it social mechanics. He published a detailed account of the new science in 1835 which he titled *A Treatise on Man and the Development of His Faculties*. This was an account of the influence of probability over human affairs.

Semiotics is the study of signs and sign processes. Charles Sanders Peirce in the nineteenth century coined the term semiotics, and he believed that signs should be used by an intelligence capable of learning by experience. This implies a set of logics be employed that govern how one operates on signs. These signs can be the concept identifiers for predicates and operated upon with predicate logics. Ferdinand de Saussure from the University of Geneva is credited with being the first to describe modern linguistics which he saw as a meaning-imbued sign.

In 1923, Ogden and Richards published *The Meaning of Meaning*. This textbook is a study of signs whose most often quoted contribution to our field is the so called Semiotic Triangle. The triangle relates thoughts, symbols, and referents by three specific relations: thought to symbol = correct, thought to referent = adequate, symbol to referent = true (see Fig. 2.2) [4]. The triangle gave a visual representation to the place

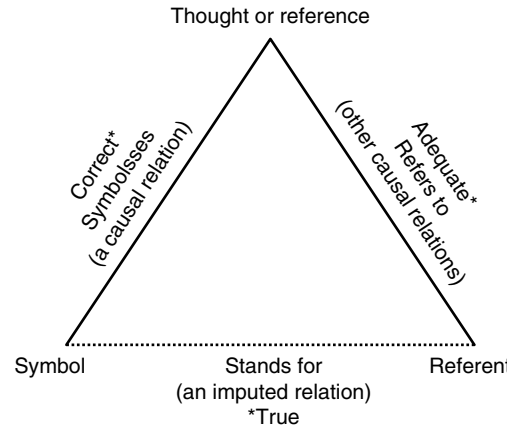


Fig. 2.2 Semiotic triangle

of symbols or in our field identifiers within the construction of a terminology or any other knowledge representation schema. This paradigm also provided a logical framework as to where language fits within our understanding of concepts. In terminological construction, the abstract thought would correspond to the thought and would not be altered regardless of the language or symbol used, and the referent is the object itself, and the symbol can be a code (identifier) or a string in a human readable language.

In 1938, Charles Morris published his seminal work dividing interoperability into three components [5]. Syntactic Interoperability deals with interoperable structures. Semantic Interoperability deals with the interoperability of a common shared meaning. Pragmatic Interoperability deals with the external constraints on the system. This last category takes into account the level of granularity needed for common understanding and the complexity or difficulty required to achieve a certain level of interoperability. Although Morris was referring to the Pragmatic Philosophers, we have extended this to address the practical side of standards development and implementation.

Chomsky published, in 1955 in mimeograph form and in press in 1975, his seminal work, *The Logical Structure of Linguistic Theory* [6]. This work expressed the view that language was a predictable, systematic, and logical cognitive activity that required a metamodel of language to effectively communicate. He demonstrated that

the behaviorists' stimulus–response model could not account for human language. This idea, that language is processed, led to the application of computer science to free text (natural language) processing. Computational linguistics (CL) is the field of computer science that seeks to understand and to represent language in an interoperable set of semantics. CL overlaps with the field of Artificial Intelligence and has been often applied to machine translation from one human language to another.

Researchers have succeeded, to varying degrees, to create CL algorithms for retrieving clinical texts. Sager in 1994 published a paper entitled, "Natural Language Processing and the Representation of Clinical Data" [7]. Here, Dr. Sager showed that for a set of discharge letters, a recall[1] of 92.5% and a precision[2] of 98.6% could be achieved for a limited set of preselected data using the parser produced by the Linguistic String Project at New York University [8–10].

Researchers have also succeeded, to varying degrees, at representing the concepts underlying clinical texts. In 1999, Wagner, Rogers, Baud, and Scherrer reported on the Natural Language generation of urologic procedures [11, 12]. Here they used a conceptual graph technique to apply translations for 172 rubrics from a common conceptual base between French, German, and English. They demonstrated that the GALEN model was capable of technically representing the concepts well; however, the language generation was often not presented in a form which native speakers of the target language would find natural. Trombert-Paviot et al. reported the results of the use of GALEN in mapping French procedures to an underlying concept representation [13]. Wroe et al. in 2001 reported the ability to integrate a separate ontology for drugs into the GALEN model [14]. Rector in his exposé *Clinical Terminology: Why Is It So Hard?* discusses the importance of and ten most challenging impediments to the development of compositional systems capable of representing the vast majority of clinical information in a comparable fashion [15]. In 2001, Professor Rector published one workable method for integrating information models and terminology models [16].

In 2004, Friedman et al. reported a method for encoding concepts from health records using the UMLS [17]. In this study, the investigators used their system, MedLEE, to abstract concepts from the record and reported a recall of 77% and a precision of 89%. In 2001, Nadkarni provided a description of the fundamental building blocks needed for NLP [18]. He discussed their method for lexical matching and part of speech tagging in discharge summaries and surgical notes. Lowe developed MicroMeSH an early MUMPS-based terminology browser which incorporated robust lexical matching routines. Lowe, working with Hersh, reported the accuracy of parsing radiology reports using the Sapphire indexing system [19]. Here, they reported good sensitivity and were able to improve performance by limiting the UMLS source vocabularies by section of the report.

Beyond representing clinical concepts, tools are needed to link text provided by clinicians to the concepts in the knowledge representation. Cooper and Miller created a set of NLP tools aimed at linking clinical text to the medical literature using the MeSH vocabulary [20]. Overall, the composite method yielded a recall of 66% and a precision of 20%. Berrios et al. reported a vector space model and a statistical method for mapping free text to a controlled health terminology [21]. Zou et al. reported a system, IndexFinder, which was principally a phrase representation system [22]. Srinivasan et al. indexed Medline citations (titles and abstracts) using the UMLS [23]. Their method took the output of a part-of-speech tagger and feeds the SPECIALIST minimal commitment parser, the lexicon used by the UMLS system. The output of this stage was matched to a set of grammars that yielded a final match.

NLM recently developed MetaMap [24]. It has the capacity to be used to code free text (natural language) to a controlled representation which can be any subset of the UMLS knowledge sources. MetaMap uses a five-step process which begins by using the SPECIALIST minimal commitment

[1]Recall = proportion of relevant texts retrieved by the algorithm. Recall is also called "sensitivity."

[2]Precision = proportion of texts retrieved by the algorithm that are relevant. Precision is also called "positive predictive value."

parser which identifies noun phrases without modifiers. The next step involves the identification of phrase variants. These variants are then used to suggest candidate phrases from within the source material [25]. Linguistic principals are used to calculate a score for each potential match. Brennon and Aronson used MetaMap to improve consumer health information retrieval for patients [26].

Elkin et al. described and validated the first practical methods for the generation of automated compositional expressions and then validated methods for terminology server creation which as of 2011 still has the highest reported accuracy in the health informatics literature. The data created using this method has been used for biosurveillance, case-based teaching, billing more accurately in healthcare, and for fully automated electronic quality monitoring (eQuality).

The history and evolution of specific terminologies and other related standards efforts will be discussed in the chapters dealing with those terminologies and their associated terminological systems. Some of the terminologies that will be discussed in this textbook are ICD, the Current Procedural Terminology (CPT) from the American Medical Association (AMA), SNOP – SNOMED – SNOMED RT – SNOMED CT, the Read Codes – Clinical Terms v2 and v3 – SNOMED CT – SNOMED CT subsets, the National Drug Formulary–Reference Terminology (NDF-RT) and RxNorm, the UMLS and its component terminologies, Logical Observations Indentifiers Names and Codes (LOINC), Medra, Medcin, First Data Bank Drug Codes, and the International Classification of Nursing Practice (ICNP) and other American Nurses Association (ANA) recognized terminologies.

Questions

1. Who developed the first health related terminology?
   (a) Heraclides
   (b) Aristotle
   (c) Plato
   (d) Hippocrates
2. Who is the father of modern medical statistics?
   (a) Aristotle
   (b) William Farr
   (c) Charles Peirce
   (d) Adolphe Quetelet
3. All are examples of Universals except?
   (a) A tree
   (b) Good
   (c) Evil
   (d) Mount Fuji
4. Which concept would be considered a particular?
   (a) Gone with the wind
   (b) A mountain
   (c) Happiness
   (d) Pain
5. Aristotle's E-type is exemplified by:
   (a) A person with blond hair.
   (b) All people are born with blue eyes.
   (c) All cars have an engine.
   (d) All people are not on Mars.
6. Aristotle's O-type is exemplified by:
   (a) Some people have blue eyes.
   (b) Some people don't like chocolate ice cream.
   (c) Some people run to stay in shape.
   (d) All of the above.
7. The precursor of the International Classification of Diseases was?
   (a) The London Bills of Mortality
   (b) The Organon
   (c) The US Classification of Diseases
   (d) The French Classification of Diseases
8. Semantics are?
   (a) The form of an expression
   (b) The terms in an expression
   (c) The concepts in an expression
   (d) The meaning of an expression
9. Semantics can be represented as:
   (a) A set of propositions
   (b) Concept and relationship/Concept triples
   (c) A syllogism
   (d) All of the above
10. Logical positivism was applied to:
    (a) Science and biology
    (b) Biology and psychology
    (c) Psychology and science
    (d) Science and biology

References

1. Bocheński IM. Ancient formal logic. Amsterdam: North-Holland Publishing Company; 1951.
2. Rose LE. Aristotle's syllogistic. Springfield: Charles C Thomas Publisher; 1968.
3. Halliday S. William Farr: campaigning statistician. J Med Biogr. 2000;8:220–7. Royal Society of Medicine Press, London, UK.
4. Richards IA, Ogden CK. The meaning of meaning. San Diego: Harvest/HBJ; 1989.
5. Charles W Morris, Foundations of the theory of signs. International encyclopedia of unified science, volumes 1 and 2 foundations of the unity of science, volume 1 number 2, Editor in chief, Otto Neurath University of Chicago Press, 1938, Ninth impression 1957.
6. Chomsky N. The logical structure of linguistic theory. New York: Plenum; 1975.
7. Sager N, Lyman M, et al. Natural language processing and the representation of clinical data. J Am Med Inform Assoc. 1994;1(2):142–60.
8. Sager N. Syntactic analysis of natural language. In: Advances in computers, vol. 8. New York: Academic Press; 1967. p. 153–88.
9. Grishman R, Sager N, Raze C, Bookchin B (1973) The linguistic string parser. In: AFIPS conference proceedings. AFIPS Press, Montvail, 1973, vol 42, pp 427–34.
10. Sager N, Gishman R. The restriction language for computer grammars of natural language. Commun ACM. 1975;18:390–400.
11. Wagner JC, Rogers JE, Baud RH, Scherrer JR. Natural language generation of surgical procedures. Int J Med Inform. 1999;53(2–3):175–92.
12. Wagner JC, Rogers JE, Baud RH, Scherrer JR. Natural language generation of surgical procedures. Medinfo. 1998;9(Pt 1):591–5.
13. Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux AM, Abrial V, Clavel L, Idir H. Galen: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. Stud Health Technol Inform. 1999;68:901–5.
14. Wroe CJ, Cimino JJ, Rector AL. Integrating existing drug formulation terminologies into an HL7 standard classification using OpenGALEN. Proc AMIA Symp. 2001;766–70.
15. Rector AL. Clinical terminology: why is it so hard? Methods Inf Med. 1999;38(4–5):239–52.
16. Rector AL. The interface between information, terminology, and inference models. Medinfo. 2001;10(Pt 1): 246–50.
17. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated Encoding of Clinical Documents Based on Natural Language Processing. JAMIA. 2004;11(5): 392–402.
18. Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. J Am Med Inform Assoc. 2001;8:80–91.
19. Huang Y, Lowe H, Hersh W. A pilot study of contextual UMLS indexing to improve the precision of concept based representation in XML-structured clinical radiology reports. J Am Med Inform Assoc. 2003; 10:580–7.
20. Cooper GF, Miller RA. An experiment comparing lexical and a statistical method for extracting MeSH terms from Clinical free text. J Am Med Inform Assoc. 1998;5:62–75.
21. Berrios DC. Automated Indexing for full text information retrieval. Proc AMIA Symp. 2000:71–5.
22. Zou Q, Chu WW, Morioka C, Leazer GH, Kangarloo H. IndexFinder: a method of extracting key concepts from clinical texts for indexing. Proc AMIA Symp. 2003:763–7.
23. Srinivasan S, Rindflesch TC, Hole WT, Aronson AR, Mork JG. Finding UMLS Metathesaurus concepts in MEDLINE. Proc AMIA Symp. 2002:727–31.
24. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NLM Indexing Initiative. Proc AMIA Symp. 2000:17–21.
25. Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17–21.
26. Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. J Biomed Inform. 2003;36(4–5):334–41.

Knowledge Representation and the Logical Basis of Ontology

3

Peter L. Elkin and Steven H. Brown

Clinical data from electronic health records have traditionally contained a small proportion of fixed field data (often obtained from pick lists) and larger quantities of free text. Some EHRs only store images of handwritten or typed notes (e.g., faxed in data). These practices have made it difficult to extract and use electronic health record data for secondary purposes.

These purposes can be categorized into assistance with the practice of medicine, research, and education. The practice of medicine can employ EHR knowledge at the point of care in the form of alerts and expert advice for the clinician, the patient, or their family (caregivers). Ideally these systems could learn from the outcomes associated with the population of patients cared for by a given provider. Research stands to gain substantially by employing EHR data for secondary uses. These can and will range from more intelligent study design where the impact on recruitment can be tested as we add additional criteria to either the inclusion or exclusion criteria for the study. We will employ data-driven recruitment that will assure that a much higher percentage of participants screened for recruitment to a clinical trial will be found to be appropriate for that trial. For retrospective trials, we will be able to run fully automated studies and complete trials in minutes rather than years. For prospective studies, we will be able to track a much broader set of clinical outcomes making more fruitful our research dollar spent. For education, real-time learning systems will be updated with the results of clinical practice and based on best outcomes will be able to educate all physicians in a practice area with information learned from anyone's practice. This continuous learning environment will advance the quality of practice available to all patients.

In order for this dream to become a reality, it requires a common data infrastructure into which all clinical data are represented. This requires defining the formalism, and then it requires a method for encoding the clinical data recorded during the normal clinical care workflow into this common representation schema. The formalism to be usable must represent the data at the same level of granularity as is recorded in routine clinical practice.

Knowledge representation is the process of designing models and systems that represent knowledge, facts, and rules.

P.L. Elkin, M.D., MACP, FACMI (✉)
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

S.H. Brown, M.D., MS, FACMI
Department of Veterans Affairs and Vanderbilt, Department of Biomedical Informatics, Nashville, TN, USA

Mapping Free Text Data into a Structured and Logical Form

Until the 1920s, logic and mathematics were often considered spiritual, not scientific. Since the time of Pythagoras, mathematics was considered a revelation of the divine order. In *Principia*

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_3,

Mathematica, Russell and Whitehead demonstrated that mathematics was logical. Logical positivism was then applied to science and psychology.

Adolphe Quetelet was one of the most influential social statisticians of the nineteenth century. His applications of statistical reasoning to social phenomena profoundly influenced the course of European social science. Quetelet had come to be known as the champion of a new science, dedicated to mapping the normal, physical, and moral characteristics. Quetelet called it social mechanics. He published a detailed account of the new science in 1835 which he titled *A Treatise on Man and the Development of His Faculties*. This was an account of the influence of probability over human affairs.

Chomsky published, in 1955 in mimeograph form and in press in 1975, his seminal work, *The Logical Structure of Linguistic Theory* [1]. This work expressed the view that language was a predictable, systematic, and logical cognitive activity that required a metamodel of language to effectively communicate. He demonstrated that the behaviorists' stimulus–response model could not account for human language. This idea that language is processed led to the application of computer science to free text (natural language) processing. Computational linguistics (CL) is the field of computer science that seeks to understand and to represent language in an interoperable set of semantics. CL overlaps with the field of Artificial Intelligence and has been often applied to machine translation from one human language to another.

Researchers have succeeded, to varying degrees, to create CL algorithms for retrieving clinical texts. Sager, in 1994, published a paper entitled "Natural Language Processing and the Representation of Clinical Data." [2] Here, Dr. Sager showed that for a set of discharge letters, a recall[1] of 92.5% and a precision[2] of 98.6% could be achieved for a limited set of preselected data using the parser produced by the Linguistic String Project at New York University [3–5].

[1]Recall = proportion of relevant texts retrieved by the algorithm. Recall is also called "sensitivity."

[2]Precision = proportion of texts retrieved by the algorithm that are relevant. Precision is also called "positive predictive value."

Researchers have also succeeded, to varying degrees, at representing the concepts underlying clinical texts. In 1999, Wagner, Rogers, Baud, and Scherrer reported on the Natural Language generation of urologic procedures [6, 7]. Here, they used a conceptual graph technique to apply translations for 172 rubrics from a common conceptual base between French, German, and English. They demonstrated that the GALEN model was capable of technically representing the concepts well; however, the language generation was often not presented in a form which native speakers of the target language would find natural. Trombert-Paviot et al. reported the results of the use of GALEN in mapping French procedures to an underlying concept representation [8]. Wroe et al., in 2001, reported the ability to integrate a separate ontology for drugs into the GALEN model [9]. Rector in his exposé "Clinical Terminology: Why is it so hard?" discusses the importance of and ten most challenging impediments to the development of compositional systems capable of representing the vast majority of clinical information in a comparable fashion [10]. In 2001, Prof. Rector published one workable method for integrating information models and terminology models [11].

In 2004, Friedman et al. reported a method for encoding concepts from health records using the Unified Medical Language System (UMLS) [12]. In this study, the investigators used their system, MedLEE, to abstract concepts from the record and reported a recall of 77% and a precision of 89%. In 2001, Nadkarni provided a description of the fundamental building blocks needed for NLP [13]. He discussed their method for lexical matching and part of speech tagging in discharge summaries and surgical notes. Lowe developed MicroMeSH, an early MUMPS based terminology browser which incorporated robust lexical matching routines. Lowe, working with Hersh, reported the accuracy of parsing radiology reports using the Sapphire indexing system [14]. Here, they reported good sensitivity and were able to improve performance by limiting the UMLS source vocabularies by section of the report.

Beyond representing clinical concepts, tools are needed to link text provided by clinicians to the concepts in the knowledge representation.

Cooper and Miller created a set of NLP tools aimed at linking clinical text to the medical literature using the MeSH vocabulary [15]. Overall, the composite method yielded a recall of 66% and a precision of 20%. Berrios et al. reported a vector space model and a statistical method for mapping free text to a controlled health terminology [16]. Zou et al. reported a system, IndexFinder, which was principally a phrase representation system [17]. Srinivasan et al. indexed Medline citations (titles and abstracts) using the UMLS [18]. Their method took the output of a part-of-speech tagger and feeds the SPECIALIST minimal commitment parser, the lexicon used by the UMLS system. The output of this stage was matched to a set of grammars that yielded a final match.

NLM recently developed MetaMap [19]. It has the capacity to be used to code free text (natural language) to a controlled representation which can be any subset of the UMLS knowledge sources. MetaMap uses a five-step process which begins by using the SPECIALIST minimal commitment parser which identifies noun phrases without modifiers. The next step involves the identification of phrase variants. These variants are then used to suggest candidate phrases from within the source material [20]. Linguistic principals are used to calculate a score for each potential match. Brennon and Aronson used MetaMap to improve consumer health information retrieval for patients [21].

Terminologies, which permit qualitative and quantitative novel term composition, have the potential to provide greater content coverage than terminologies which are restricted to the use of precoordinated terms [22–24]. Postcoordination of terms can move our field considerably closer to the ultimate goal of content as well as knowledge completeness and consistency. Postcoordinated compositional terminologies can be expressively powerful, but may carry the risk of generating expressions whose equivalency cannot easily be determined [25, 26]. In order for terminologies to provide comparable data they must be normalized with respect to both their content and semantics. Without a powerful knowledge representation format that does not limit comparisons to purely qualitative forms but to model-theoretic quantitative forms as well, the potential benefits of a controlled vocabulary can become a veritable quagmire where clinicians end up building tomorrow's legacies today [27]. Precoordinated compositional expressions should retain the same representational structure as postcoordinated expressions [28]. A limited semantic model will lead to a loss of information (lossy "knowledge compression") in the postcoordinated compositional expressions. Essentially, what is required is a postcoordination with the effect of lossless "knowledge compression" where a complex postcoordinated expression may be faithfully expanded from a reduced form or compared to other expressions with ease and with no loss of semantic integrity.

Clinically useful controlled vocabularies must represent healthcare concepts completely and with high reliability and minimal redundancy [29]. The core of a medical terminological service also forms the basis of a high-quality resource of deep medical knowledge whose value over time can greatly increase over and above its initial inception [24]. However, anticipating and precoordinating all possible expressions (e.g., "fracture of the left femur" and "fracture of the right femur") may not be feasible nor useful in time-critical or research-critical scenarios where the access to the knowledge is of greater importance than the access to just the data or the information alone. Reasons include variance in practice styles, content complexity, exponential growth of terminology size, and increased terminology maintenance costs, as well as the return on value from just accumulation of more "data" versus increasing the depth of knowledge [25].

Compositional terminologies [26] are one potential solution to the problem of content completeness, but carry a risk of generating expressions whose equivalency cannot be easily determined. Several ISO standards are currently being provided more as guidelines from "lessons learned" than actual providing "the right way" [16–18] with more in development on the horizon as lessons learned in other efforts have shed light on the crucial importance of choosing a good knowledge representation format [27]. In order for postcoordinated expressions to be comparable, a formal terminological composition and decomposition mechanism for lossless

information representation is necessary, such that both explicit and implicit information can be faithfully recovered, maintained, modified demand and with ease [28]. In order to define comparison functions for postcoordinated expressions, normalization of both the contents and the semantics of the contents of the terminology needs to be accomplished by an appropriate knowledge representation format that facilitates the process. In addition, the representation format that permits such a functional comparison requires a storage and messaging paradigm robust enough to represent completely the explicit or implicit information contained within arbitrarily complex compositional expressions. We present a formalism for storing, comparing, and sending messages containing compositional expressions using a large-scale reference terminology.

Sowa et al. have described formalisms for the semantics and grammars contained in free text [30]. In medicine, we have the added advantage of concept level understanding of a large proportion of the noun phrases, which occur in medical text [31]. McGuinness et al. have described the use of subsumption in description logics to provide a basis for logical inferencing [14]. This feature of description logics has often been the basis for their usefulness as the infrastructure of reference terminologies, but has also been a source of severe troubles [27, 32]. By creating formal compositional definitions for terms in a vocabulary, one can automatically classify all fully defined concepts and place them into the correct location within a hierarchy. Further one can check for conflicts (no two concepts should have the same formal definition, unless they are synonymous). The set of health concepts is clearly not closed, but in the context of a single version of a reference terminology, a domain of the set can appear to be closed (Scott Topology) [15]. One question is whether or not there is a function on the knowledge representation such that the distance between two concepts within and between hierarchies in a terminology can be usefully quantified and that takes account of contextual uses. We suggest a method for measuring relative contextually based conceptual distance, which we name "conceptual relativity."

We recognize that relative distance measures among different metrics can lead to different relative groupings of concepts. In that sense, there are no "correct" or "incorrect" metrics. The correctness of a metric will depend on how well it meets a specific quantitative requirement. Furthermore, we propose an approach by which formal distance functions (i.e., metrics) can be worked out in order to handle appropriately the different categories of partial matches. It is incumbent upon the terminology community to continue to work toward handling approximate matches intelligently [33–35]. The field of controlled health terminologies has been plagued by a lack of having a unifying model that addresses all problems, for example, functional proximity (things are "close together" because they belong to the same functionality), data-oriented proximity (things are "close together" because they belong to the same types and attributes characterizing the data), or time-oriented proximity (things are "close together" because they happen within a short time interval of each other).

The vertical dimension in a terminology relates terms to their meanings (via specified relationships), and the horizontal dimension provides the comparative links between meanings and their expressions.

We suggest several requirements in building a compositional terminology, without repeating previous requirements that are already well known [36, 37]:

1. That the terminology is composable on both the symbolic (language term) as well as the continuous (numerical system) level so that metrics can be devised and applied.
2. That the system has a theory of *compositional knowledge* so that new knowledge can be added at any time. This implies a representation scheme that is "model-theoretic" and that various theories such as "topology" or theory of wholes and connectivity, "morphology" or theory of form and congruence, "mereology" or theory of parts, "mereotopology" the theory of connected parts, as well as "temporal mode theories" theory of time representation, and others can be added into the meta-level system without severely impacting the object-level system (of terminology).

3. The system must have a clear Continuous Terminological Semantic Theory. This will explain how both humans and machines are able to extract meaning from the controlled vocabulary. Lexical (terminological) semantics aims to decipher two things – (a) how meaning can be extracted from novel term compositions and (b) what the nature is of the meanings of the smallest meaningful units of the vocabulary – and our constraint that the terminology is "continuous" means that new compositions can be created to extend or fill in "gaps" in knowledge (or work with partial knowledge) such that a metric is available to measure or compare terms.
4. That composition and decomposition of terms are symmetric and without knowledge loss (i.e., that the operations in taking apart a term faithfully recover those elements that put it together in the first place).
5. That the controlled vocabulary and the terminological representation are compatible, upward from the state-of-the-art commercial databases available (i.e., today's OODBMS systems).

There are a variety of styles and methods by which concepts and knowledge are discussed, shared, and disseminated such that a great deal of meaning is obfuscated within specialized contexts or intents in discourse. This is particularly evident when we examine our own metalanguage (i.e., the way researchers discuss our own field in the literature). This clearly can lead to inconsistencies in the way that we design and implement controlled health terminologies. This manuscript defines some basic concepts in our field in terms of logic and focuses, in particular, on a representation independent method of adding a continuous domain representation to the symbolic or terminological representation. This method and approach of communication is unambiguous and provides the sort of rigor that our field needs to move forward in a unified direction.

However, while the focus will be on the metric for conceptual relativity, a few characteristics that will influence the initial metric spaces and forms as well as to situate the reader on familiar ground will be developed herein. All relationships have rules with which they should be applied. They provide the basis for recognizing when data are not well formed, which can serve to curtail applications from performing inappropriate inferencing. The relation between a source type and its elementary target type can be asymmetric [38], and by analogy, so should the distance metric be asymmetric, if appropriate. The same is true for the transitive closure of this relation. For example, the child of a parent cannot have the parent as a child of itself [39]. Another important generic invariant rule is that for two component instances in an ordered compositional association, it is possible to define whether one is before the other [40, 41]. For exhaustively enumerated subtype relations to be considered a surrogate definition for a supertype, the union of sets of instances of all subtypes must be equal to the elements of the set of the supertype.[xi]

Description logics form the cardinalities of the relationships used in compositional expressions. These relations have a "mandatory participation" within an association, if the existence of an instance of the entity implies the existence of a corresponding instance of the association. For example, entities within an anatomy hierarchy will always have topography [11].

Conceptual graphs form another competing representation scheme to DL in medical terminological systems [23, 33, 34]. A *canonical basis* [39] is the set of conceptual graphs encoding the elementary relationships of a domain (somewhat similar to the T-Box of DL). The canonical basis allows more composition and decomposition by applying *canonical composition rules (and reversing the rules for decomposition).*

Conceptual graphs (CGs) have three current advantages over DL:

1. It is easy to use methods of spectral graph theory to the CGs because they are graphs.
2. CGs subsume DLs as specialization in particular.

The CG form of DL T-Boxes is a restricted subset of canonical graphs which can be modeled as order-sorted feature types; the canonical formation rules on these graphs are covered by feature term unification; and the generation of the closure of a canonical basis is a special application of dependency grammar, which ties CGs closely to natural language [42].

3. The mapping of the canonical conceptual graphs presented in to the feature types of LOGIN, a feature-based extension of PROLOG presented in, was almost one-to-one. This b implementation of canonical conceptual graphs in a commercial logic programming language provided immediate operational semantics, and today, the same principles with more modern logical language systems (such as ECLIPSE-PROLOG, SICSTUS-PROLOG, or PROLOGIX-PROLOG™) can aid implementers in entering a model-test-model cycle of terminology design to avoid potential design flaws the vocabulary engineering process.

This chapter will introduce some logical notation and the basic methodology in accordance with Wells and Bagchi [43, 44].

Symbolic Logic

Glossary

Expression	Meaning
$\forall x$	For all x
$\exists y$	There exists a y
$a \supset b$	If a then b
$a \subset b$	a is a type of b
$a \wedge b$	a and b (a intersection b)
$a \vee b$	a or b (a union b)
$a \bullet b$	a concatenated with b
$\sim a$	Not a
$x \in y$	x is a member of y
$x \equiv y$	x is true if and only if y is true
$x = y$	x is an equivalent statement to (a synonym of) y

Identity of Compositional Expressions Can Be Expressed in the Following Fashion

$a \bullet a = a$; a concatenated with itself is itself.

A terminology is a finite set of words pertaining to the objects in a specific domain of discourse. The elements of a terminology are called terms. For example, the following set is a terminology: {Diseases, Cardiovascular Diseases, Congestive Heart Failure, LV Systolic Dysfunction, LV Systolic Dysfunction, Class III HF}

A terminology is not enough since there must also be relationships that determine how one term stands relatively to another term to produce an ordered structure.

When the relationships are parameterized by a distance metric, then the distance metric provides the implicit discriminatory measure by which the terminological compositions can be quantitatively evaluated. Therefore, specific relationships have specific metrics, and therefore, metrics scale according to the type of the relationship between terms. For example, dog as pet versus bear as pet compared to dog as mammal versus bear as mammal.

There are two important assumptions in this method. The first is that the terminology has normalization of both the content and semantics of the terminology. Both lexical and semantic closures are required. In lay terms, this would mean that for every concept in the terminology, one could identify mathematically all the different representational forms that can have the same meaning. These requirements may not be valid in presently available terminologies. Therefore we propose using the compositional expressions directly to look for equivalence and that the distance function of the composition provides the quantitative validation of equivalency.

There are two principal considerations in building a controlled terminological vocabulary: one is categorization of terms, as in the construction of concepts into hierarchies and the other is context. A distance metric will need to account for both relative differences in category as well as context. Then we shall provide the general form of a metric of conceptual relativity that accommodates indexing and retrieval in a representation-free manner (as "Conceptual Relativity") according to relatively scaling metrics based on a novel application of distance field methods to create a terminological semantic field that exhibits several desirable properties and advantages over previous methods. Most mathematical procedures used to create metric spaces are a combination of vector space modeling techniques coupled with a variant of ordination such as cluster analysis, principal components analysis, or multidimensional scaling. Vector space methods condense a data set of terms into a reduced-dimension distance-metric matrix typically characterized

by term cooccurrences. A major drawback is that all semantic (intrinsic) properties of data are collapsed into the extrinsic property of a numerical value, usually based on Euclidean distance. Functional or temporal "distance" as well as other types of scaled metrics disappears and multiple levels of hierarchies, feature structures, and mereotopological (regional or topographic) details are not preserved. The ability to have a comparative data representation that is preserved at different levels of granularity is therefore lost, and this prevents effective compositional value from being fully realized.

Fig. 3.1 Objects have multiple valid perspectives from which they can be viewed

Fundamentals of Object-Oriented Modeling

- Abstraction
- Modeling
- Structured Analysis
- Object Orientation
- Software Robustness

Abstraction

- Looking only at the information that is relevant at the time
- Hiding details so as not to confuse the bigger picture

Modeling

- Each view has information that is unique to that view (see Fig. 3.1).
- Each view has information that appears in at least one other view.
- The information that is common between views is consistent.

Structured Analysis (see Fig. 3.2)

- Functional view
- Data view
- Dynamic view

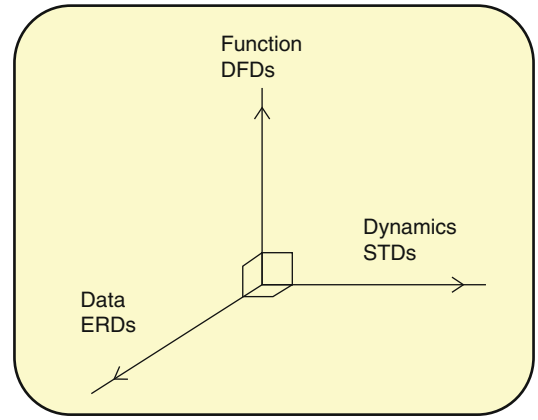


Fig. 3.2 The multiple perspectives of structured analysis

Object Orientation

Object orientation includes a class model, an interaction model, and a dynamic model (see Fig. 3.3).

What Is UML?

UML (see Fig. 3.4) is a well-structured and non-proprietary language. It provides a set of notations and rules for specifying a software design. UML primarily focuses on creating simple, well-documented, and easy to understand software designs.

- Simple and structured
- Language, process, and tool independent
- Architecture – centric

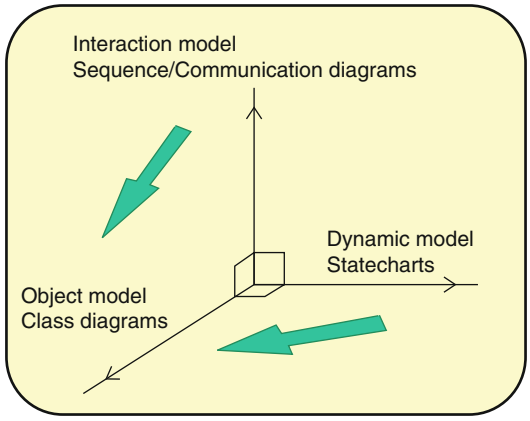


Fig. 3.3 Object orientation requires a formal analysis against a set of dimensions including the interaction model, the dynamic model, and the object model

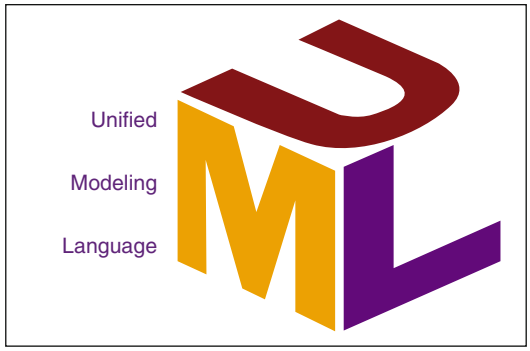


Fig. 3.4 UML logo

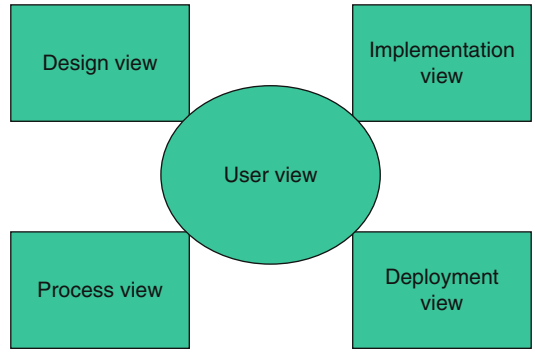


Fig. 3.5 Multiple consistent views of the UML model

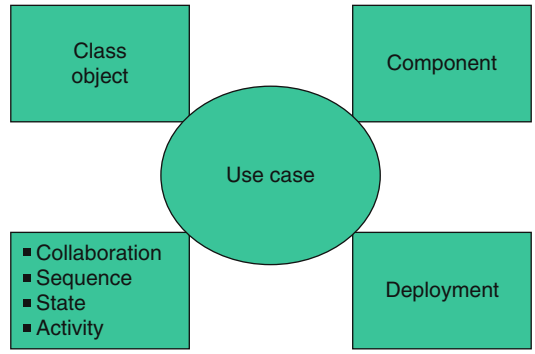


Fig. 3.6 Use cases influence the UML objects

UML Views

The user view interacts with a design view, the implementation view, the process view, or the deployment view (see Fig. 3.5).

UML Diagrams

UML diagrams are use case driven, and the components take their context from the use cases for the model (see Fig. 3.6).

Class Diagram

Class diagrams define the objects, their relationships to other objects or actors, and the cardinalities associated with these relationships (see Fig. 3.7).

Class Attributes

The attributes of a class define the class and are both instance and class data members (see Fig. 3.8).

Class Operations

The operations of a class combine the methods associated with the class along with the arguments for that method and the return types of the objects returned by that method (see Fig. 3.9).

Visibility

- Use visibility markers to signify who can access the information contained within a class (see Fig. 3.10). Private visibility hides information from anything outside the class partition. Public visibility allows all other classes to view the marked information. Protected visibility allows child classes to access information they inherited from a parent class.
- Visibility is the logical visibility of the UML element, which is similar to how "public" is used in a program – not how the node is displayed.

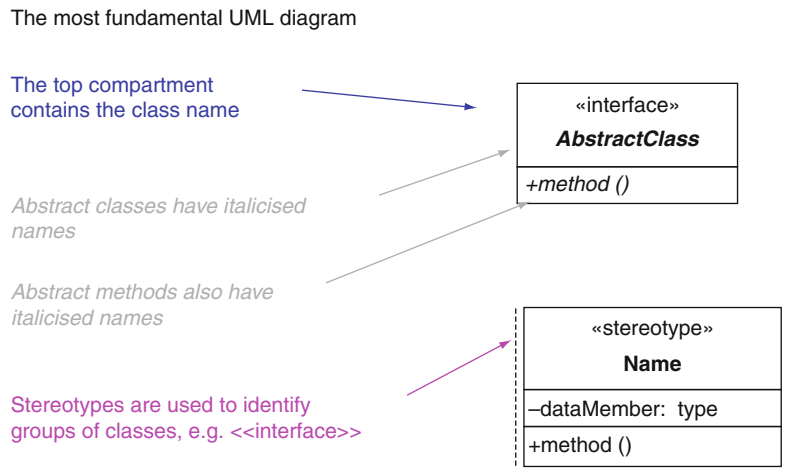


Fig. 3.7 UML diagrams define classes of knowledge

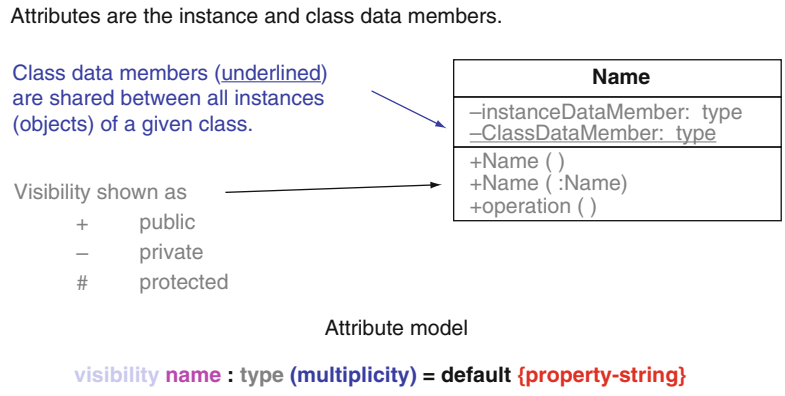


Fig. 3.8 Class attributes for the defining features of a class

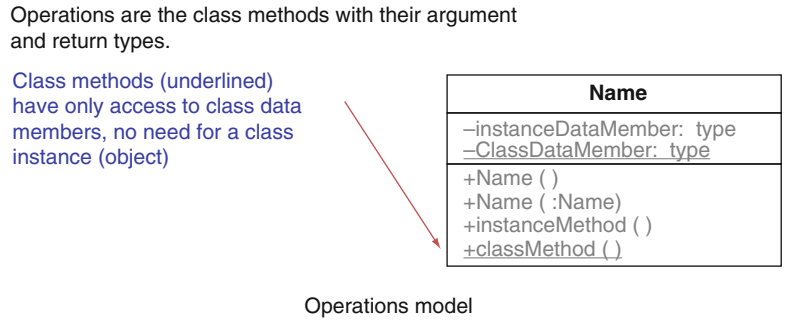


Fig. 3.9 Class operations and their methods, arguments, and return types

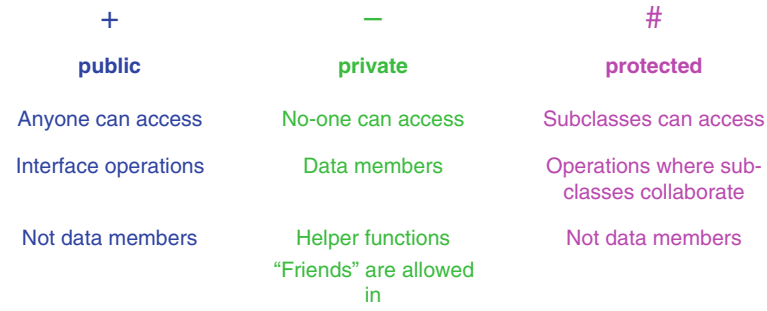


Fig. 3.10 The features of public, private, and protected classes

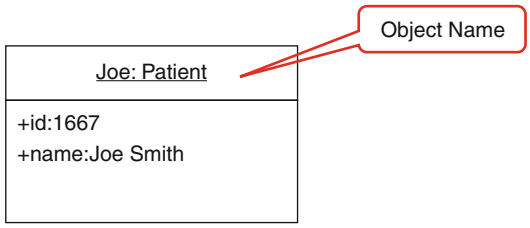


Fig. 3.11 Object Joe is in context patient and has an ID and a name

Object Diagram

An object diagram (see Fig. 3.11) defines the object and often relates that information to a class within a class diagram.

Use Case Diagram

A use case diagram shows the functionality of the system from an outside-in viewpoint (see Fig. 3.12).

Sequence Diagram

Sequence diagram shows potential interactions between objects in the system being defined (see Fig. 3.13).

Collaboration Diagram

Collaboration diagram shows similar information to sequence diagrams but with a different perspective (see Fig. 3.14).

Class Model for Database Design

The class definitions in the model help us to define the structure of the database tables and the behavior of the database (see Fig. 3.15).

Mapping Tables to the Classes

Tables and their definitions can be used to develop class diagrams (see Fig. 3.16). This has to be done with care so that one does not create incompatible constructs across the various implementations within a system as the context may vary across implementations. The context-based definitions must remain constant across implementations.

Mapping Columns to the Attributes

Often the attributes should be used to formally define the column names within a UML representation or its associated database design (see Fig. 3.17).

Views

Virtual representation of data collected from multiple tables. Views are logical but not physical table structures (see Fig. 3.18). However, one can operate on a view just like it was an actual table.

Keys

Public key and foreign keys define unique identifiers of a table and normalized links from one table to another in the database (see Fig. 3.19). These serve as constraints on the data. For example, if you have a column for patient in the physician table, the patient as determined by their unique identifier (ID) may be constrained to exist in the patient table.

Constraints

A constraint is a rule applied to a column, table, or schema. Here, the product type serves as a constraint (see Fig. 3.20).

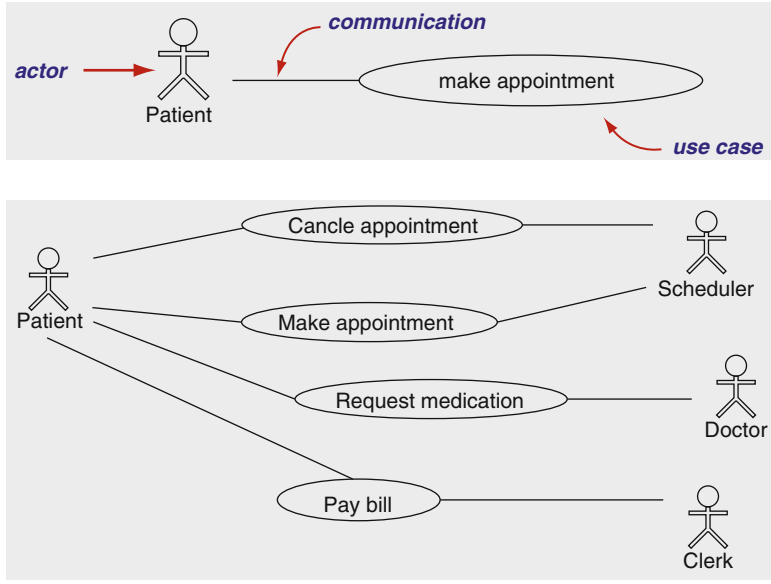


Fig. 3.12 Example of a clinical use case diagram where a patient makes an appointment or can cancel an appointment and can request a medication from their physician or pay a bill to the hospital

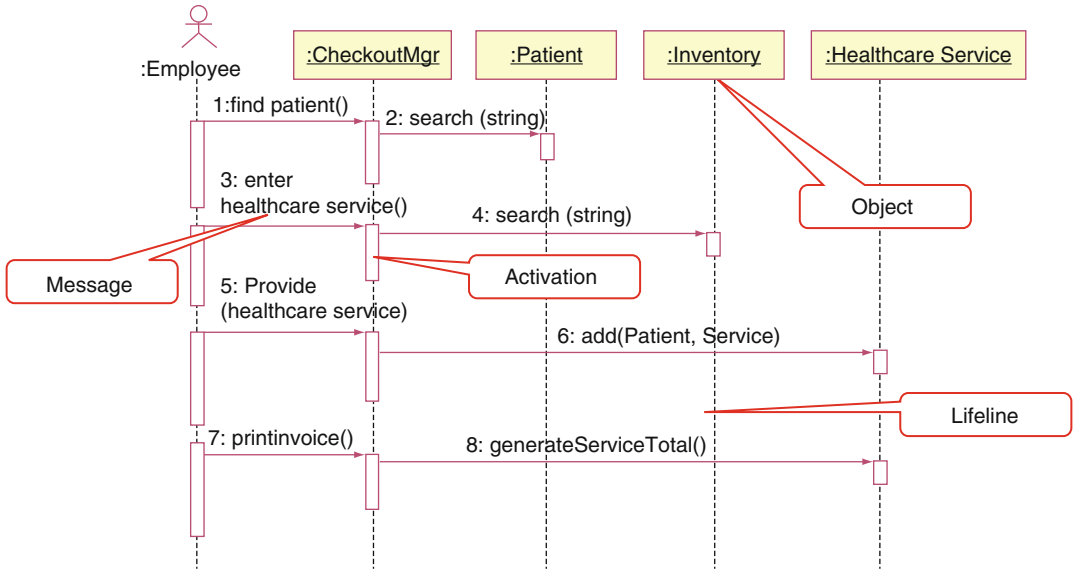


Fig. 3.13 Sequence diagram showing a patient receiving a healthcare service

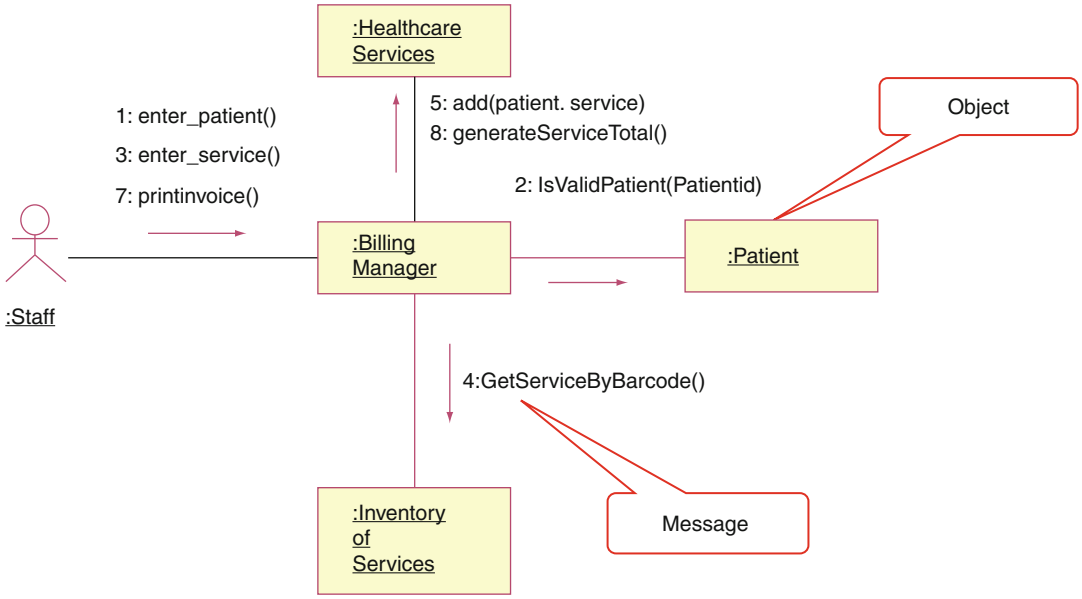


Fig. 3.14 Collaboration diagram showing how the actors in the model interact

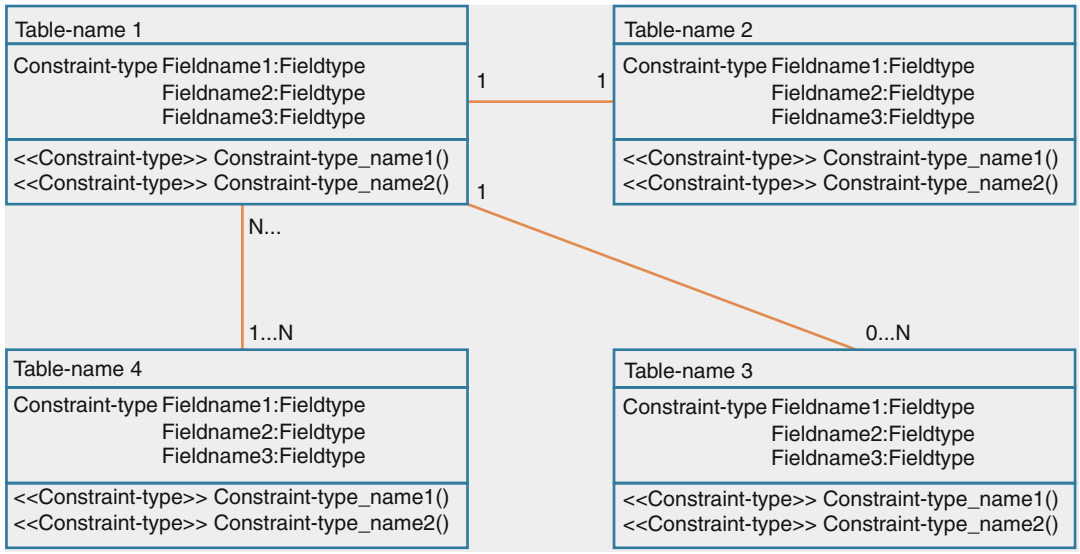


Fig. 3.15 Class models influence the database structure

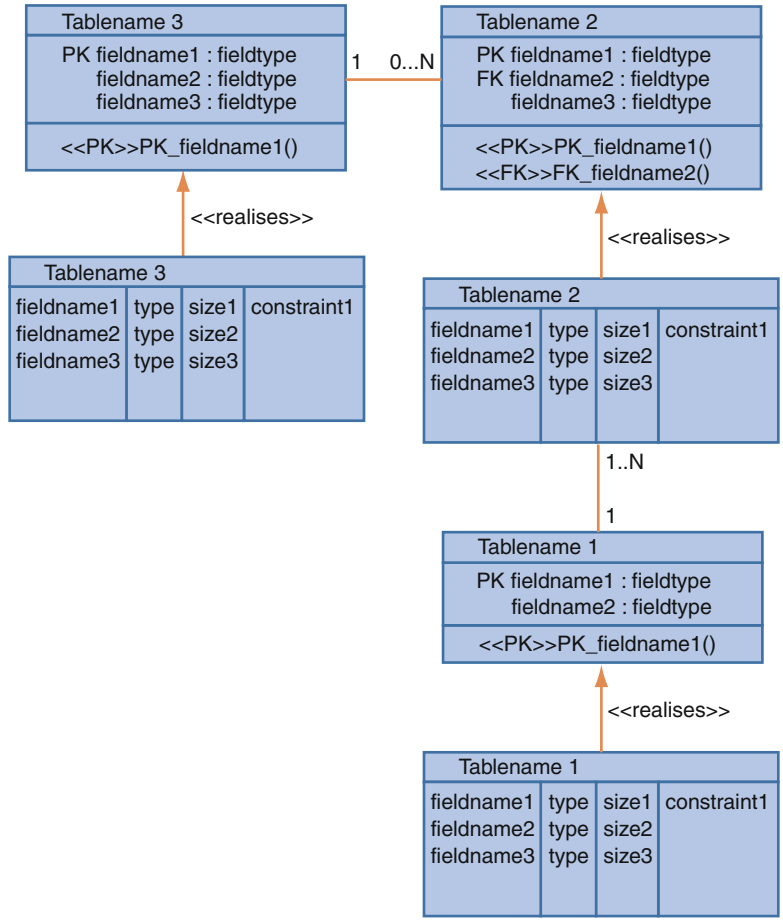


Fig. 3.16 Mapping tables to classes requires context to prevent ambiguity

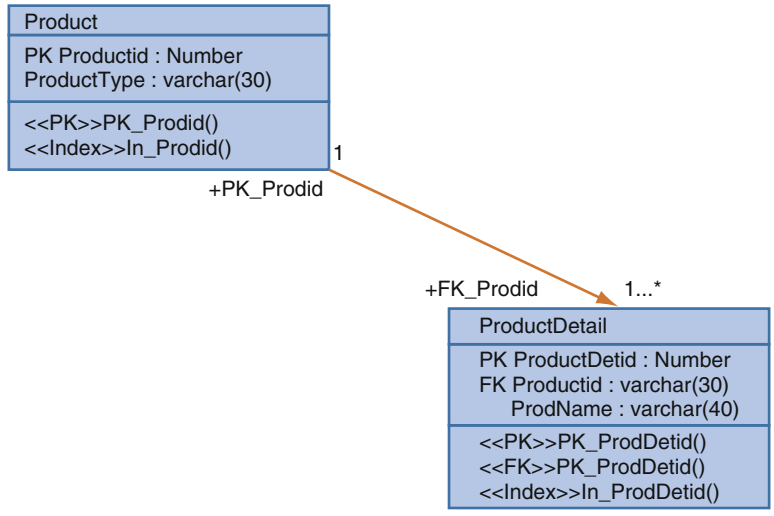


Fig. 3.17 Mapping columns to attributes serves to define the column names

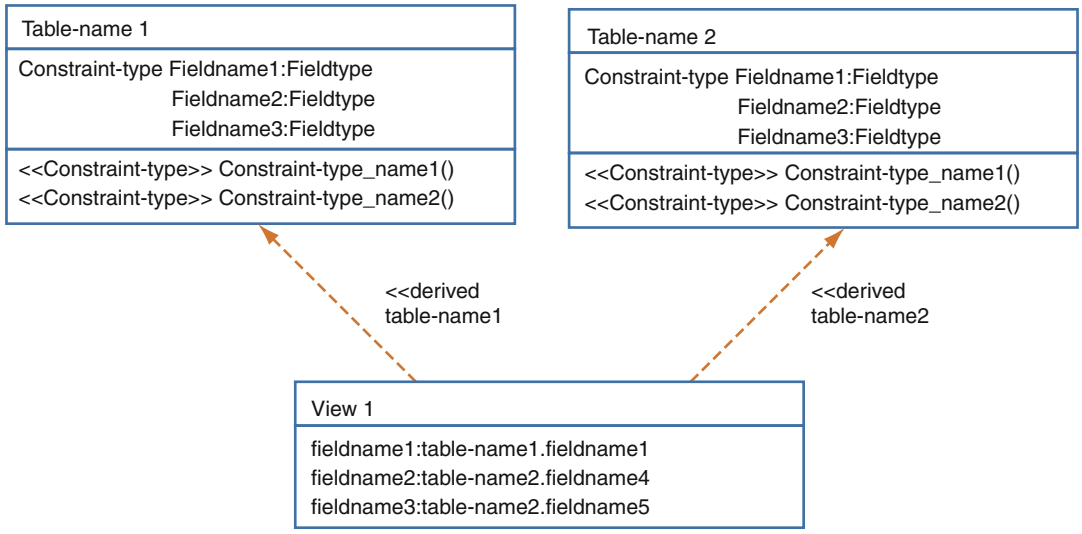


Fig. 3.18 Views are a logical representation of a table without their needing to be a physical representation of that table

<<PK>>, <<FK>>, <<PFK>>
Public key, Foreign key, Public / Foreign key

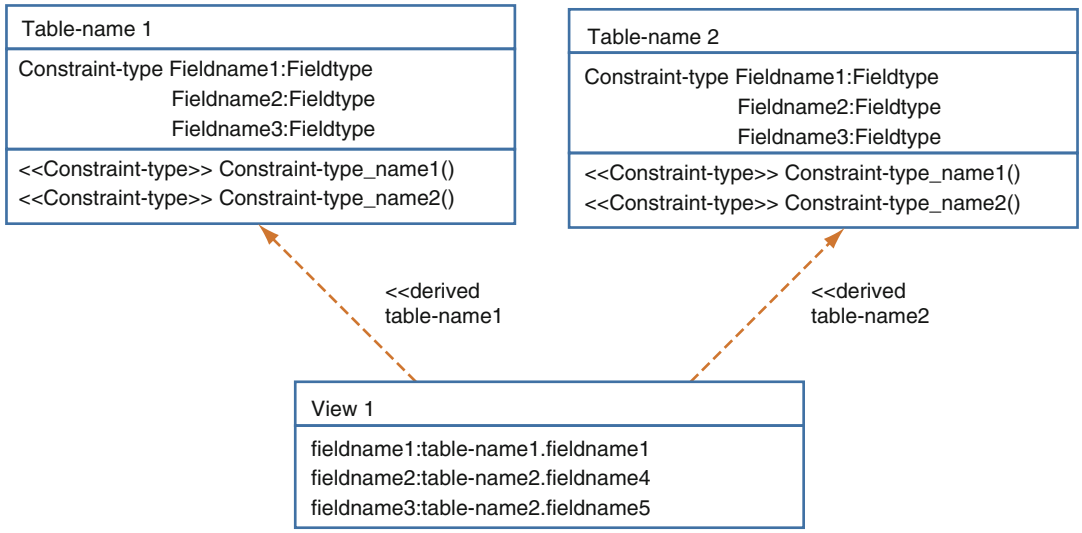


Fig. 3.19 Public Key Infrastructure (PKI) shows the relationship and use of public and private keys for data security

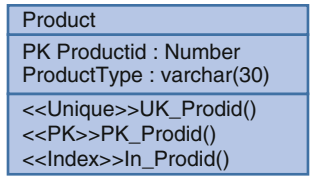


Fig. 3.20 Constraint on the type of product as an example of the application of a constraint. This product has a specific type

Object Constraint Language (OCL)

The Object Constraint Language is a notational language for analysis and design of software systems. It is a subset of the UML that allows software developers to write constraints and queries over object models.

UML diagrams are limited as they do not describe constraints about objects or classes. Also natural language can be ambiguous, and designers can specify unambiguous rules using OCL constructs.

OCL can be used:

- To specify *invariants* on classes and types in the class model
- To specify a type *invariant* for stereotypes (areas of the model reused by reference {e.g., linkage of phenotype to sequence or microarray data})
- To describe *pre-* and *post*conditions on operations and methods
- To specify constraints on operations

Invariant

An invariant is something that must be true for all instances of the class or its subclasses (see Fig. 3.21). This represents the transitive reflexive closure of subsumption. Where the transitivity is reflective of a property being inherited by all its subtypes, and the reflexive property means that the name of the type of classes is itself a member of the class (e.g., asthma is a type of asthma without having to specify it specifically as a member of the class of asthmas which include, but are not limited to, exercise-induced asthma, hypersensitivity asthma, etc.). An OCL expression is an invariant of the classifier and therefore must be true for all instances of that classifier at any time. (Note that all OCL expressions that express invariants are of the type Boolean.)

Invariant in the context of the healthcare organization class

Hospital.numberOfEmployees > 50

Here *Hospital* is an instance of type Healthcare Organization. This invariant holds for every instance of the Healthcare Organization type.

Fig. 3.21 As an example of an invariant constraint, a hospital must have >50 employees

Pre- and Postconditions

The OCL expression can be part of a precondition or postcondition, corresponding to «precondition» and «postcondition» stereotypes of constraint associated with an operation or method (see Fig. 3.22).

OCL Types

- Predefined types
  - Basic. types – integer, real, string, and Boolean
  - Collection types – set, bag, sequence
    - Set – A mathematical set, elements are unique and not ordered.
    - Bag – A group of elements, but may contain duplicates.
    - Sequence – A bag with elements ordered.
- Meta types
  - oclType, oclAny, oclExpression
    - *oclType* – instance for all the types defined in a UML model or predefined within OCL. Hence, each OCL type is an object instantiated from oclType type (see Fig. 3.23).
    - *oclAny* – a supertype of all types in the model and the basic predefined OCL

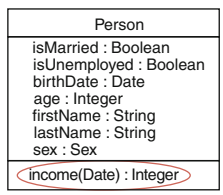


Fig. 3.22 Example of preconditions and postconditions in the context of a person

Pre- and post conditions in the context of the Person::income method

context Person::income(d: Date) : Integer
pre: d > 12.31.2000
post: result = 5000

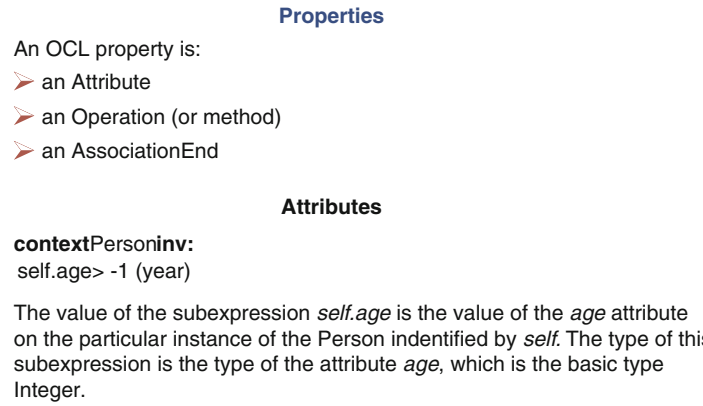


Fig. 3.23 OCL type showing properties and attributes relative to a person

type. The predefined OCL collection types are not subtypes of oclAny.
    - *oclExpression* – a type for each OCL expression.
- User-defined model types
  - All classes, types, and interfaces from UML diagrams (UML classifiers).

Collection Operations

OCL defines many operations on the collection types. These operations are meant to project new collections from existing ones.

In most of the cases the multiplicity of an association is not 1, but more than 1. Evaluating a constraint in these cases will result in a collection of instances of the associated class.

Constraints can be put on either the collection itself, e.g., limiting the size, or on the elements of the collection.

- Select and Reject
  - Special constructs to specify a selection from a specific collection.
- Collect
  - Specifies a collection that is derived from some other collection, but which contains different objects from the original collection.
- ForAll
  - Constraint on all the elements.
- Exists
  - Allows a specification of a Boolean expression that must hold for at least one object in a collection.
- Iterate
  - Very generic operation, which accumulates new collection by iterating over another collection. The operations *reject*, *select*, *forAll*, *exists*, and *collect* can all be described in terms of *iterate*.

Object Management Group's Unified Modeling Language version 2

What Is a Model?

- Something that accurately resembles something else; a thing that represents on a small scale the structure or qualities of something greater [OED 2002] http://www.omg.org/spec/UML/2.0/.
- A model captures a view of a physical system. It is an abstraction of the physical system, with a certain purpose. This purpose determines what is to be included in the model and what is irrelevant. Thus the model completely describes those aspects of the physical system that are relevant to the purpose of the model at the appropriate level of detail [UML 2001].
- In the UML metamodel, model contains a containment hierarchy of model elements

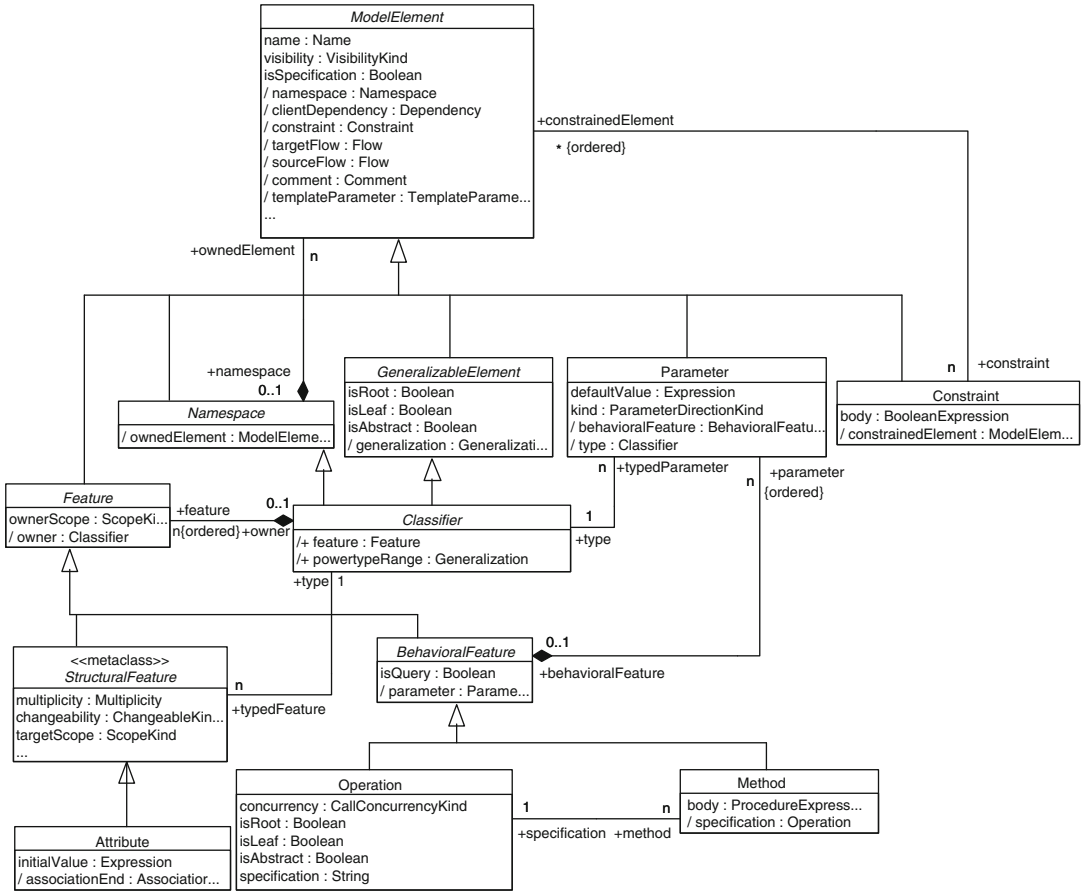


Fig. 3.24 UML foundation core class diagram which provides context for all UML models

that together describe the physical system. A model also contains a set of model elements that represents the environment of the system, typically actors, together with their interrelationships, such as dependencies, generalizations, and constraints.

- Different models can be defined for the same physical system, where each model represents a view of the physical system defined by its purpose and abstraction level (e.g., an analysis model, a design model, an implementation model).
- Typically different models are complementary and defined from the perspectives (viewpoints) of different system stakeholders. For example, a use-case model may be defined from the viewpoint of a business analyst stakeholder.

UML Model: Foundation Core

The UML class model for the foundation core provides the context for all UML models (see Fig. 3.24). Model elements are parts of the model and can contain classifiers to help to determine the correct logical assignments for new classes (see Fig. 3.25).

UML Metamodel: Model Management

The management of UML models is governed by the UML metamodel model (see Fig. 3.26). This may seem to be overspecified; however, as models become large, consistency in their usage protects against procedural ambiguity. This

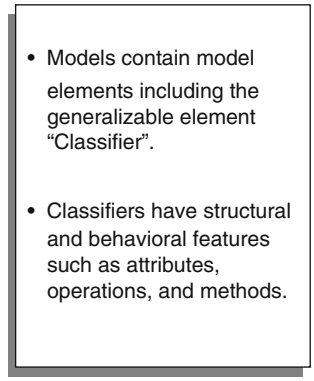


Fig. 3.25 Classifier elements can have attributes, operations, and methods

would allow changes in meaning with different methods of use, for example, use in different namespaces which have different definitions. Packages group models that are to be referenced together (see Fig. 3.27).

UML Metamodel: Common Behaviors

The common behaviors of the metamodel show the actions available to UML models

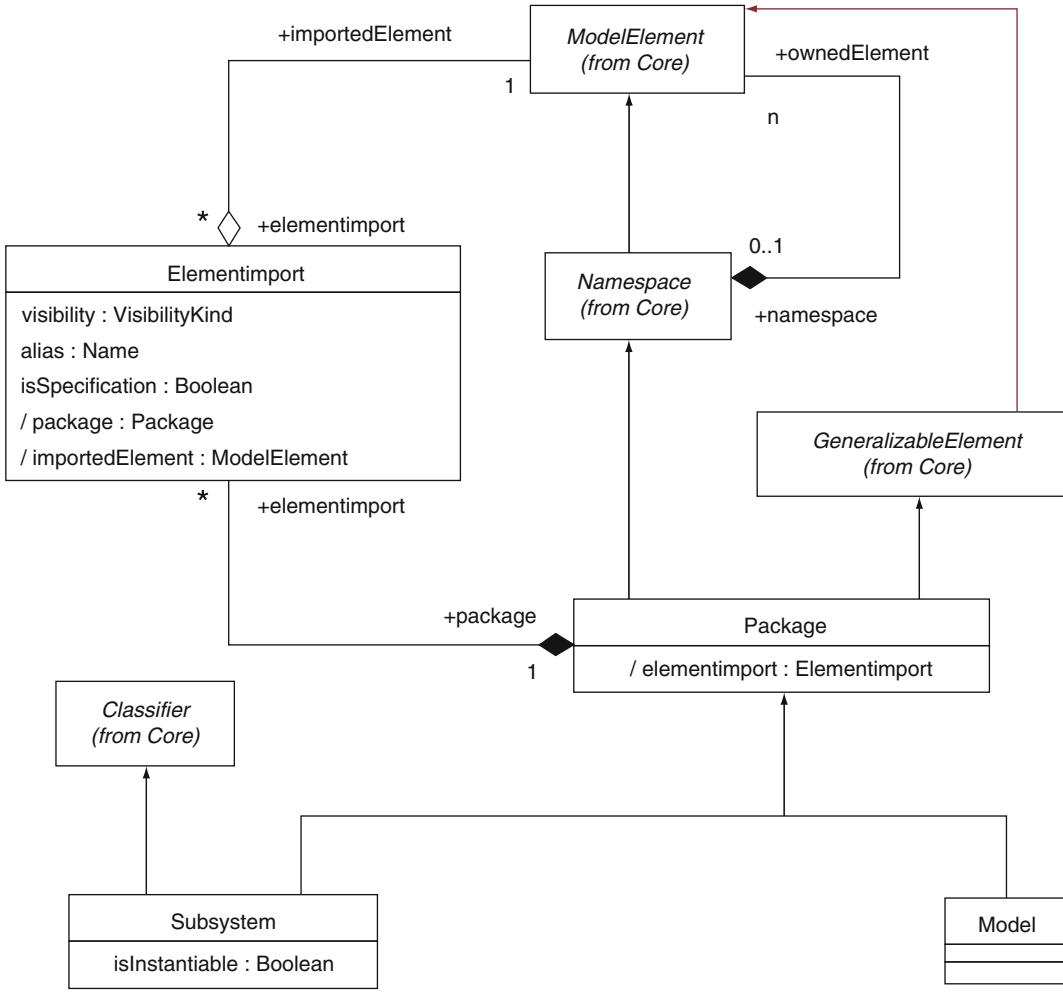


Fig. 3.26 UML metamodel which specifies how models are to be used

(see Fig. 3.28). This includes how propositions are constructed and constrained. This includes the definition of actions and an ordered set of arguments (see Fig. 3.29).

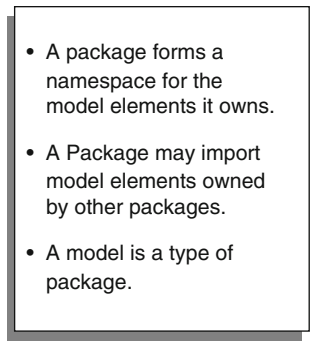


Fig. 3.27 Relationship between packages and models and namespaces

UML Metamodel: Extension Mechanisms

UML extension mechanisms include (see Figs. 3.30 and 3.31):

- Stereotype
  A stereotype is, in effect, a subclass of an existing metamodel element with the same form (attributes and relationships) but with different intent. A stereotyped element may have additional constraints on it from the base metamodel class. It may also have tagged values that add information needed by elements branded with the stereotype.
- Tag definition
  Tag definitions specify new kinds of properties that may be attached to model elements. The actual properties of individual model

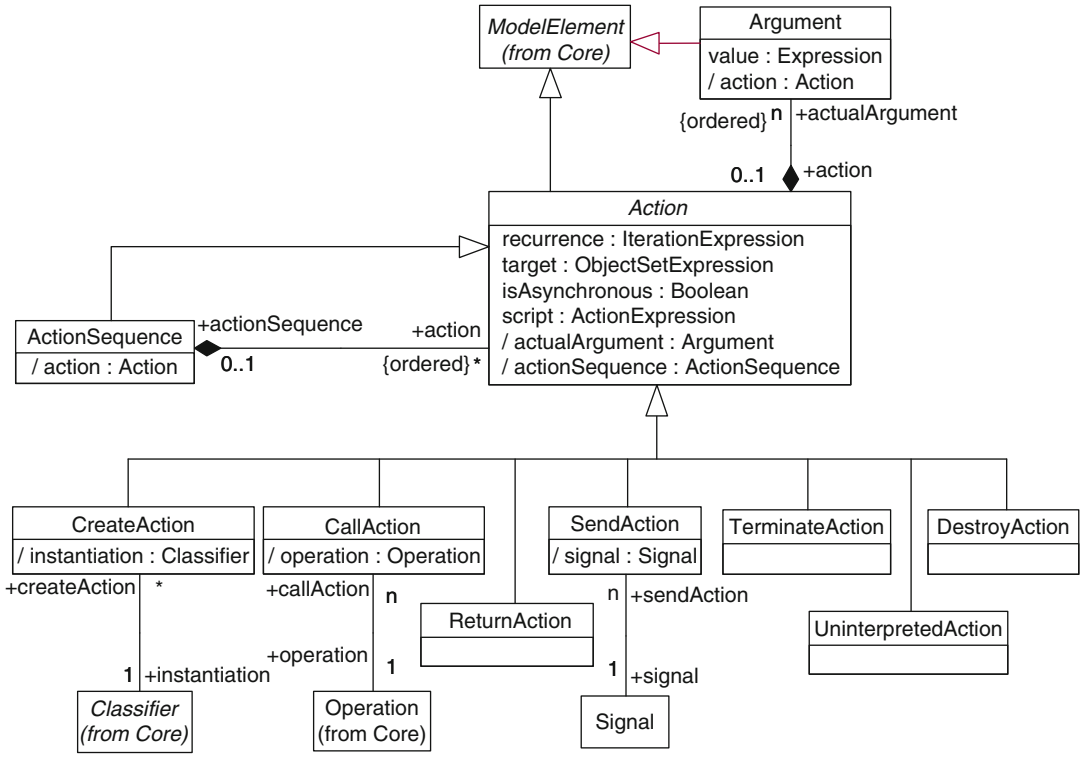


Fig. 3.28 UML metamodel of propositions and actions

elements are specified using tagged values. Tag definitions are used to define the virtual metaattributes of the stereotype to which they are attached.

- Stereotype constraint
  Designates constraints that apply to all model elements branded by the stereotype to which they are attached. A constraint is semantic information attached to a model element that specifies conditions and propositions that must be maintained as true; otherwise, the associated model element is not well-formed.
- Tagged value
  A tagged value is a keyword–value pair that may be attached to any kind of model element. The keyword is called a tag. Each tag represents a particular kind of property applicable to one or many kinds of model elements.

Web Ontology Language (OWL)

- "The OWL Web Ontology Language is a language for defining and instantiating *Web ontologies*."

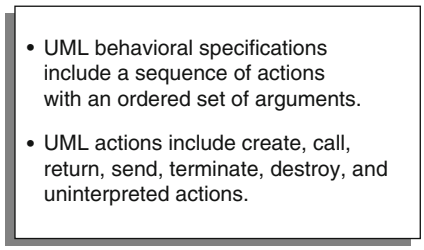


Fig. 3.29 UML behavioral specifications

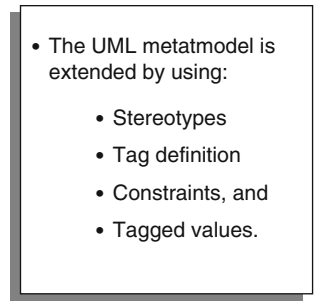


Fig. 3.31 Description of UML extension mechanisms

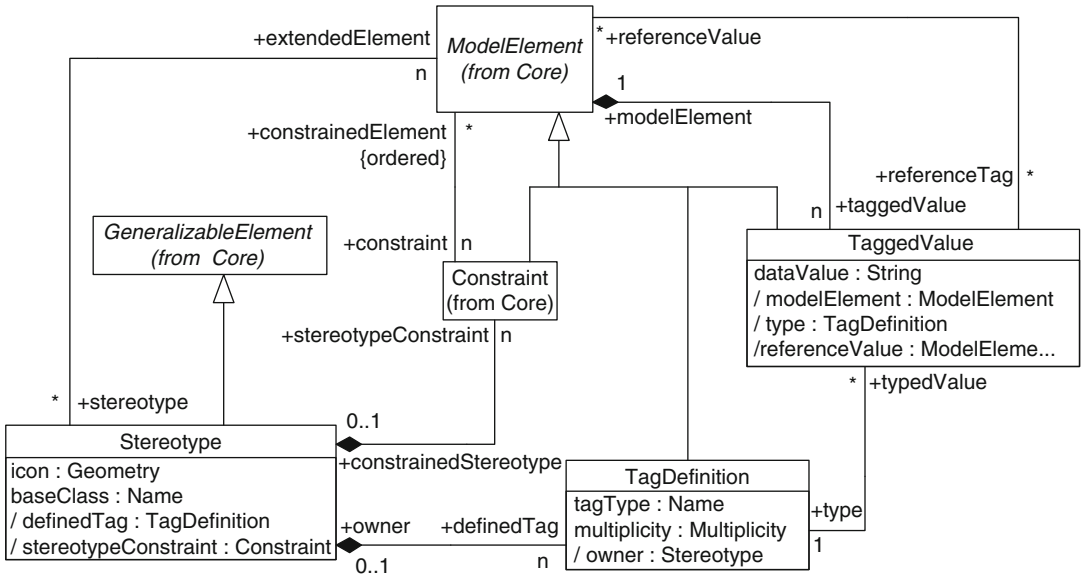


Fig. 3.30 UML metamodel of extension mechanisms

- Developed by the W3C Web Ontology Working Group (WebOnt), in support of the Symantic Web.
- Based on DAML+OIL.
- RDF is the syntax for OWL.
- OWL is a formalism that provides for the description of classes, properties, and instances of the same.
- OWL formal semantics can be applied against OWL ontologies to reason out facts that are contained within the ontology, but may not be explicitly defined.
- In a nutshell: To enable computers to *understand* the semantic content of documents.

Ontology Definitions

<Philosophical>

A systematic account of existence.

<Artificial intelligence>

An explicit formal specification of how to represent the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them.

<Information science>

The hierarchical structuring of knowledge about things by subcategorizing them according to their essential (or at least relevant and/or cognitive) qualities.

Ontology: An organization of concepts for which one can make a rational argument. Colloquially this term is used to describe a hierarchy constructed for a specific purpose. For example, a hierarchy of qualifiers would be a qualifier ontology.

What Is the Semantic Web?

> The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation
>
> Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, *Scientific American*, May 2001.

Use of Ontologies in Health Informatics

- Level One Ontologies
  - Domain independent
  - Things that are true about the whole world
  - Example, HL7 RIM
- Level Two Ontologies
  - Domain dependent
  - Things that are true for a particular domain
  - EHR architecture
- Level Three Ontologies
  - Nomenclatures
  - Rules for the formation of compositional expressions
  - Archetypes

An example of a level one ontology in healthcare is the HL7 Reference Information Model (see Fig. 3.32). It talks about entities which have

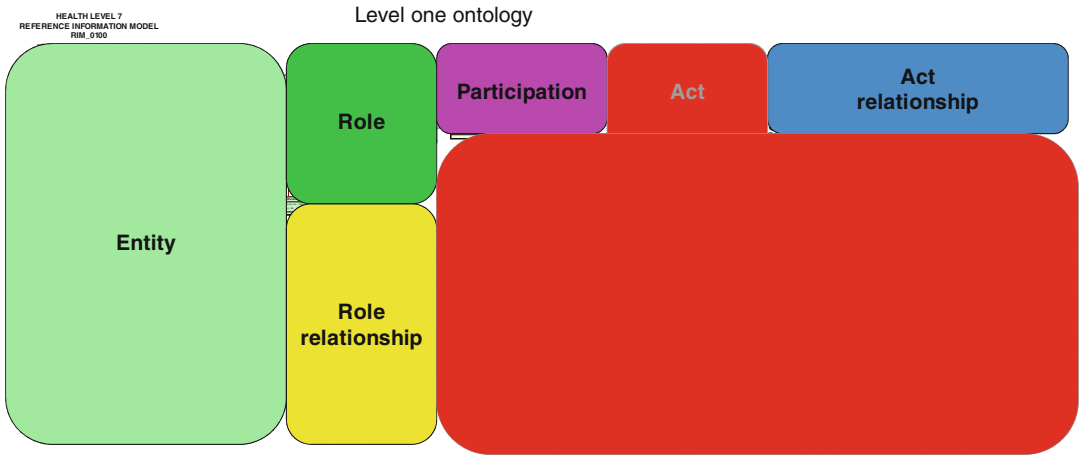


Fig. 3.32 HL7 RIM as an example of a level one ontology

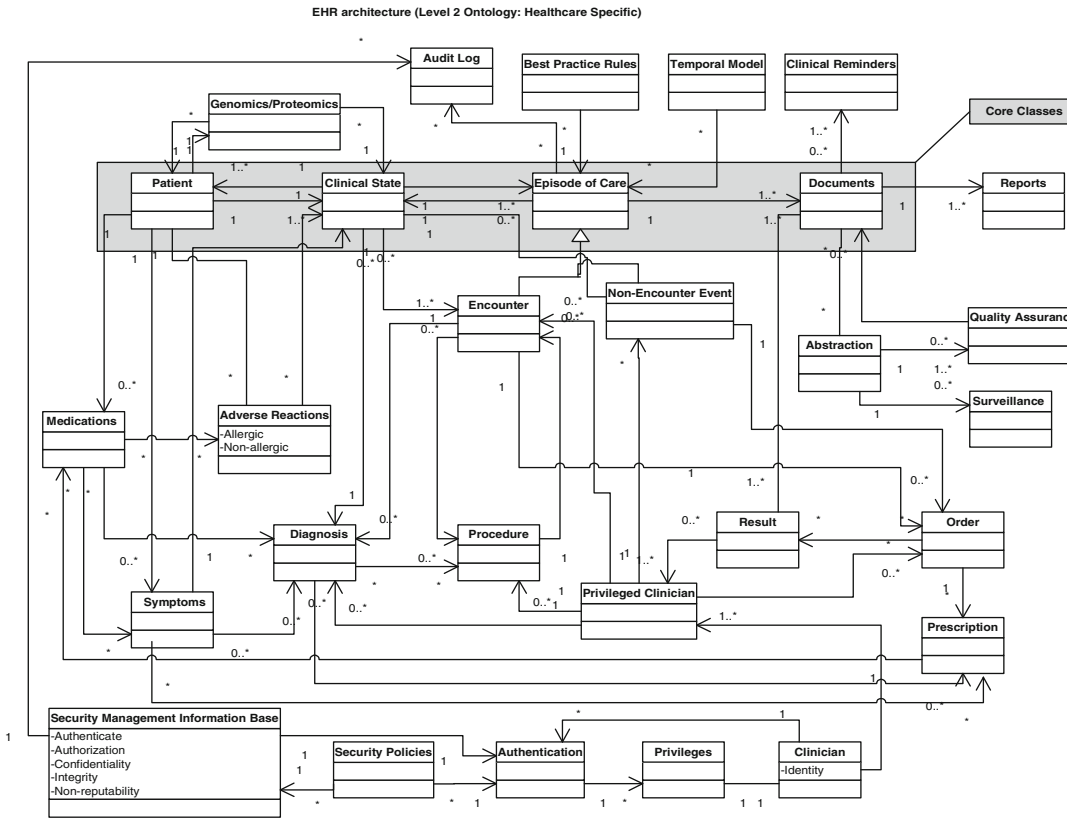


Fig. 3.33 An EHR UML model designed by our team as an example of a level two domain ontology

roles that may have a role relationship which can have a participation in some act which can have one or more act relationships. This is domain independent and as such represents things that are true about the world in general.

The level two ontology is a domain ontology (see Fig. 3.33). Here, we represent knowledge related to a specific field or discipline. This model has to be consistent with the level one ontology as it inherits its properties and restrictions. The level two ontology constrains the level one ontology by specifying knowledge as it relates to the domain of interest, in our case, health and healthcare. Here, we describe classes such as medications, diagnoses, procedures, and all information goes through a privileged clinician, and these events aggregate into an episode of care which serves as a single point of audit. The model describes clinical states separately from the patient as they are frozen in time and represent a step toward the four-dimensional health record. Further this design is a fully secure health record and incorporates the ideals and content of the ASTM and ISO security standards.

The level three ontology is the fully encoded medical record. This is the details describing what can go wrong with the patient and is often best shown as the instance data for a particular patient's health record often referred to as the fully encoded health record (see Fig. 3.34). It must be consistent with the level one and two ontologies for health. Its compositional expressions are assigned *automagically*. It was once said that technology sufficiently advanced appears to be magic. Here, the automatic assignment of canonical codes and a single logical representation is an important part of our efforts toward interoperable healthcare data. It is important that no human effort is needed to construct such a representation as this would create an undue clinical and economic burden on the prac-

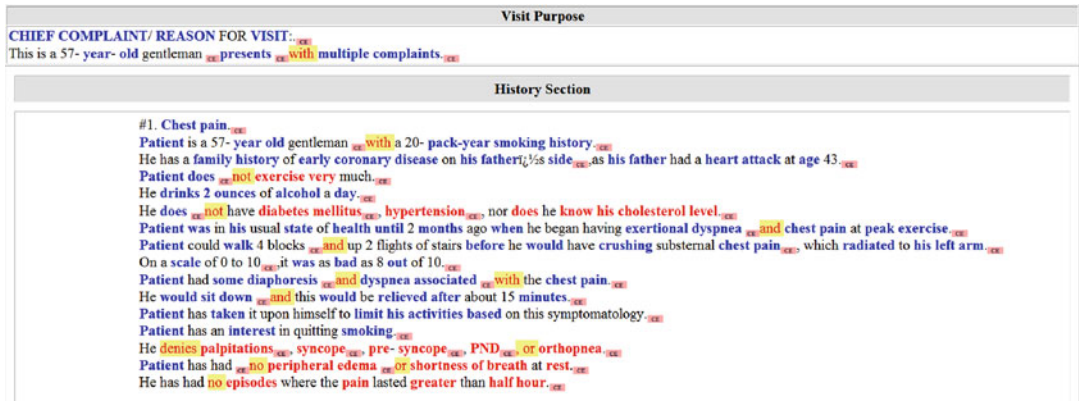


Fig. 3.34 A level three ontology or the fully encoded health record. Here, the concepts in color are codified and linked in this example to SNOMED CT codes. The *blue* concepts are positive assertions, the *red* are negative assertions, and the *green* are uncertain assertions

tice of medicine. The practice of medicine should not have to be shaped around information systems but instead the information systems can unobtrusively be employed to improve patient care.

This commonality of meaning can lead to multicenter data sharing and therefore can facilitate broad-based clinical trials, sharing quality and safety rules, biosurveillance, and other practice improvement programs.

Formal definitions are a type of knowledge representation. Knowledge representation is the application of logic and ontology to the task of constructing computable models for some domain. Formal definitions are composed of ontological entities and logical statements in a way that can be reduced to symbols and manipulated algorithmically. Ontology defines the kinds of things that can exist in the application domain. Logic provides formal structure and the rules of inference. Formal definitions allow use of computer-based terminology tools to check for duplicate definitions; to check for logically inconsistent definitions and to algorithmically assign class memberships to individuals rather than require human curators to assign class membership explicitly.

There are different types of formal logic that differ in their expressiveness and theoretical computability. Propositional logic, first order logics and higher order logics represent a spectrum of increasing expressiveness and complexity. Many logics used in computer science, including the description logics that are commonly used in formal terminologies, are subsets of first order logics (see Fig. 3.35).

Resource Description Framework (RDF)

This is a language written to communicate information with definitions within the World Wide Web and is a standard of the W3C. RDF supports application processing of web information (as opposed to human readability.) It maintains the semantic meaning of information as it is communicated between intelligent agents. RDF identifies resources through URIs. It is an XML-based syntax, whose model is a set of RDF triples.

There are three parts to an RDF triple:

1. Subject, an RDF URI reference or a blank node
2. Predicate, an RDF URI reference
3. Object, an RDF URI reference, a literal or a blank node

The predicate is referred to as the property of the triple.

Web Ontology Language (OWL)

This also is a standard of the W3C and is a formal description logic-based language used to represent knowledge on the Internet. OWL is the Web Ontology Language and is the ontological representation language for the Semantic Web of the

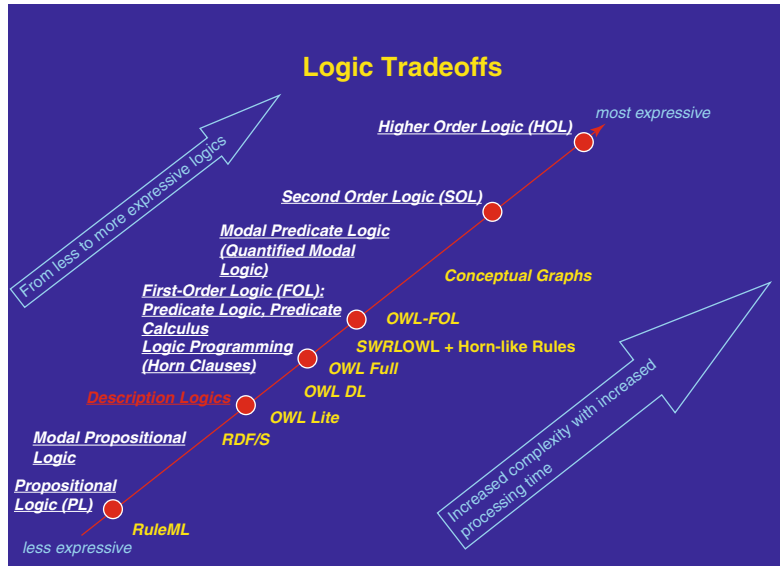


Fig. 3.35 Expressivity vs. complexity of logic systems

W3C. OWL was developed by the W3C Web Ontology Working Group (WebOnt) in support of the Semantic Web. "The OWL Web Ontology Language is a language for defining and instantiating *Web ontologies*." OWL is based on DAML + OIL which were created by the military researchers and Stanford University. On the web, RDF and RDF schema are the syntax for OWL.

There are many other knowledge representation languages including KL-one, K-rep, FACT, Ontolog, KRSS, Protégé, DAML, OIL, and many others. In this chapter, we will focus on OWL as it has the widest usage and will serve as an example for most of the functions of the other languages.

XML or the extensible markup language is the web standard within which RDF and RDFS:OWL are written. By itself, XML only provides the capability to create human readable definitions. In order to hold and use large ontologies, we need to be able to specify computable definitions which are definitions that can be understood and operated upon by computers. We need to be able to autoclassify the ontology. Description logics ask the logical question is B subsumed by A, where A and B are classes of information. This is a very powerful question and allows us to know based on a concept formal definition where it belongs in the ontology and if there are any other concepts with the same formal definition. This is what we mean by autoclassify or to automatically classify the location of all parts of the ontology. This also allows us to ensure that there is no ambiguity of the information meaning that there is no concept with more than one meaning and no redundancy meaning that there are not two concepts with the same meaning. This is important for interoperability of healthcare data. Lastly, this is necessary for us to be able to connect the information model and the terminological model.

OWL is a formalism that provides for the description of classes, properties, and instances of the same. OWL formal semantics can be applied against OWL ontologies to reason out facts that are contained within the ontology, but may not be explicitly defined.

There are three sublanguages (or species) of OWL (see Fig. 3.35):

- OWL Lite
  - For rapid translation of taxonomies
  - Limited restriction set
- OWL DL (description logic)
  - Contains the full set of OWL language constructs, but places limits on the use of the constructs
  - Ensures that computations will complete in real time

- OWL Full
  - The big kahuna.
  - Provides full expressiveness of RDF.
  - "It is unlikely that any reasoning software will be able to support every feature of OWL Full." (W3C)

OWL Lite provides support for building a classification hierarchy. It provides the basic constraint model (i.e., cardinality is limited to values of 0 or 1) of the OWL language. OWL like constructs should be relatively easy to craft supporting tools or integrate with a description logic reasoner. OWL Lite was primarily developed as a tool for developers, with same semantic restrictions as OWL DL, and was aimed as an easy method to transfer terminologies into a web ontology.

OWL DL is a proper sublanguage of *OWL Full.* OWL DL computations will complete in a finite period of time. This language contains the full set of OWL constructs, but places limits on the use of the constructs. Principally OWL DL does not allow the same name to be used for more than one type of construct including an object, a datatype, an object property, or a datatype property. This prevents complex higher order logical constructs with complexities that cannot be computed in any reasonable period of time (e.g., the lifetime of a human being).

OWL DL requires a pairwise separation between classes, datatypes, datatype properties, object properties, annotation properties, ontology properties (i.e., the import and versioning stuff), individuals, data values, and the built-in vocabulary. This means that, for example, a class cannot be at the same time an individual.

In OWL DL, the set of object properties and datatype properties are disjoint. This implies that the following four property characteristics:

inverse of,
inverse functional,
symmetric, and
transitive

can never be specified for datatype properties

OWL Full is the most expressive and flexible of the OWL sublanguages. It can handle the highest complexity. It is not required for most knowledge representation tasks. For example, OWL Full is not required for designers to build an ontology. OWL Full is more difficult to use and more easy to get into trouble with by creating either highly complex or nonlogical constructs. "OWL Full allows free mixing of OWL with RDF Schema and, like RDF Schema, does not enforce a strict separation of classes, properties, individuals and data values." "OWL DL puts constraints on the mixing with RDF and requires disjointness of classes, properties, individuals and data values."

Each of these sublanguages is an extension of its simpler predecessor, both in what can be legally expressed and in what can be validly concluded. The following set of relations hold; however, their converses do not hold as true:

- Every legal OWL Lite ontology is a legal OWL DL ontology.
- Every legal OWL DL ontology is a legal OWL Full ontology.
- Every valid OWL Lite conclusion is a valid OWL DL conclusion.
- Every valid OWL DL conclusion is a valid OWL Full conclusion.

Basic OWL Elements

- Classes
- Individuals
- Properties
  - Datatype properties
  - Object properties

OWL classes define a classification for an *individual* or instantiation of a *class.* OWL classes group concepts with similar attributes together into a *class* of objects. *Individual* members of a *class* inherit the properties of the same *class.* In OWL Lite and OWL DL, owl:Class (or owl:Restrictions, see further) must be used for all class descriptions. owl:Class is defined as a subclass of rdfs:Class. The rationale for having a separate OWL class construct lies in the restrictions on OWL DL (and thus also on OWL Lite), which imply that not all RDFS classes are legal OWL DL classes. In OWL Full, these restrictions do not exist, and therefore owl:Class and rdfs:Class are equivalent in OWL Full.

Class descriptions allow designers of OWL ontologies to craft computable definitions for the

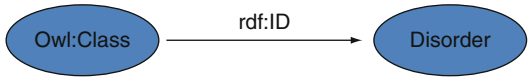


Fig. 3.36 An RDF triple in the namespace ex: describing the relationship that disorders come from a class defined by an RDF:ID or a specific type of concept such as the SNOMED CT disorder hierarchy

classes belonging to the ontology. Classes can be defined either directly (*class identifier*) or based on the description of the attributes of their members (*class descriptor*). There are six types of c*lass* descriptions that either describe a class by a name or a constraint:

1. A class identifier
2. Enumeration of individuals
3. Property restrictions
4. Union
5. Intersection
6. Compliment

Class Identifiers

A class identifier can be used to define a class through the class name.

In OWL syntax, this is accomplished by declaring an RDF URI reference (universal index of registered name spaces and addresses.)

<owl:Class rdf:ID="Disorder"/>

This asserts the RDF triple:

ex: (see Fig. 3.36)

where ex: is the namespace for the ontology being referenced.

Enumeration

Utilizes the built in *oneOf* property

An enumerated list of the individuals (or instances) is used to define the class.

The below example identifies the class of the types of disorders.

```
<owl:Class>
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="#Cardiovascular
    Disorders"/>
    <owl:Thing rdf:about="#Gastrointestinal
    Disorders"/>
    <owl:Thing rdf:about="#Rheumatological
    Disorders"/>
    <owl:Thing rdf:about="#Hematological
    – Oncological Disorders"/>
    <owl:Thing rdf:about="#Neurological
    Disorders"/>
        …..
  </owl:oneOf>
</owl:Class>
```

Property Restrictions

- Defines a class of all individuals (instances) that meet the restriction criteria
- Two types of property restrictions
  - Value
    - Constrains the class based on the range of the property
  - Cardinality
    - Constrains the class based on the number of properties

Value Restrictions

- Possible value constraints include:
  - owl:allValuesFrom
    - For each instance of class being described, all properties constrained by the class definition must be present.
    - A simple example constraining a class myocardial infarction to be a cardiovascular disorder that takes its values from SNOMED CT.
  - owl:someValuesFrom
    - For each instance of class being described, at least one of the properties constrained by the class definition must be present.
  - owl:hasValue
    - For each instance of class being described, the properties constrained by the class definition must have the declared value.

```
<owl:Class rdf:ID="Myocardial Infarction">
  <rdfs:subClassOfrdf:resource="&Disorders;
  Cardiovascular Disorders "/>
      …
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource =
        "#hasMorphology"/>
```

```
            <owl:hasValue rdf:resource=
            "#SNOMED CT:Infarction"/>
          </owl:Restriction>
          <owl:Restriction>
            <owl:onProperty rdf:resource=
            "#hasLocation"/>
            <owl:hasValue rdf:resource=
            "#SNOMED CT:Myocardium"/>
          </owl:Restriction>
        </rdfs:subClassOf>
          …
  </owl:Class>
```

If we wanted to specify that we will get all our values from SNOMED CT that would look like:

```
  <owl:Class rdf:ID="Myocardial Infarction">
    <rdfs:subClassOf rdf:resource=
    "&Disorders; Cardiovascular Disorders "/>
          …
      <rdfs:subClassOf>
        <owl:Restriction>
          <owl:onProperty rdf:resource=
          "#hasMorphology"/>
          <owl:AllValuesFrom rdf:resource=
          "#SNOMED CT"/>
        </owl:Restriction>
        <owl:Restriction>
          <owl:onProperty rdf:resource=
          "#hasLocation"/>
          <owl: AllValuesFrom rdf:resource=
          "#SNOMED CT"/>
        </owl:Restriction>
      </rdfs:subClassOf>
          …
  </owl:Class>
```

If we wanted to say that the values could come from either SNOMED CT or ICD9-CM, we would write:

```
  <owl:Class rdf:ID="Myocardial Infarction">
    <rdfs:subClassOf rdf:resource=
    "&Disorders; Cardiovascular Disorders "/>
          …
      <rdfs:subClassOf>
        <owl:Restriction>
          <owl:onProperty rdf:resource=
          "#hasMorphology"/>
          <owl:SomeValuesFrom
          rdf:resource= "#SNOMED CT"/>
        </owl:Restriction>
        <owl:Restriction>
          <owl:onProperty
          rdf:resource="#hasMorphology"/>
          <owl: SomeValuesFrom
          rdf:resource="#ICD9-CM"/>
        </owl:Restriction>
      </rdfs:subClassOf>
          …
  </owl:Class>
```

Cardinality Constraints

Possible cardinality constraints include:

- owl:cardinality
  - Specifies the exact cardinality for the property being constrained
  - NOTE: OWL Lite is limited to cardinality values of 0 or 1
- owl:maxCardinality
  - Specifies the maximum cardinality for the property being constrained
- owl:minCardinality
  - Specifies the minimum cardinality for the property being constrained

A simple example of a cardinality constraint is Age.

Here, we are constraining the class of Age that has resource Years:

```
  <owl:Class rdf:ID="Age">
          …
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty
        rdf:resource="#hasYears"/>
        <owl:maxcardinality rdf:datatype="
          &xsd;nonNegativeInteger">140</
          owl:maxcardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
          …
  </owl:Class>
```

Property Constraints

Provides a mechanism for property restrictions based upon set operations

OWL-defined constraints include:

- owl:unionOf
  - Constrains the value of a class to be the union of two individuals

- owl:intersectionOf
  - Constrains the value of a class to be the intersection of two individuals
- owl:complementOf
  - Constrains the value of a class to be the complement of an individual
  - NOTE: owl:complementOf is not allowed in OWL Lite

We define the set of nonoperative procedures as:

```
<owl:Class>
  <owl:Restriction>
  <owl:onProperty rdf:resource=
  "#hasProcedure"/>
  <owl:AllValuesFrom rdf:resource=
  "#SNOMED CT:Procedure"/>
  </owl:Restriction>
<owl:complementOf>
  <owl:Class rdf:about=
  "#OperativeProcedure"/>
  </owl:complementOf>
</owl:Class>
```

where operations and nonoperative procedures are subtypes of class of procedure.

OWL Individuals

- *Individuals* are described as the members (or instances) of a particular *class.*
  - Individual members are different from member variables in object-oriented programming languages such as Java.
- Individuals are instantiated by specifying facts (called axioms).
- There are two types of axioms used to declare individuals:
  - Axioms asserting class membership and property values for individuals
  - Axioms asserting individual identity

```
<CardiovascularDisorderrdf:ID="Myocardial
Infarction"/>
```

In the example above, the disorder myocardial infarction (more commonly known as a heart attack) inherits all of the properties of a cardiovascular disorder. The statement asserts myocardial infarction to be a member of the class of cardiovascular disorders.

Individuals can also be instantiated by asserting individual identity.

OWL has three built-in constructs for asserting individual identity.

- owl:sameAs
  - Asserts the fact (or axiom) that two URI references reference the same individual
- owl:differentFrom
  - Asserts the fact (or axiom) that two URI references reference different individuals
- owl:AllDifferent
  - Asserts that the listed individuals are all different

Owl:differntFrom example:

```
<Disorder df:ID="Cardiovascular
Disorder">
  <owl:differentFrom rdf:resource="#Gastr
  ointestinal Disorder"/>
  <owl:differentFrom
  rdf:resource="#Orthopedic Disorder"/>
  <owl:differentFrom rdf:resource="#Opht
  halmological Disorder"/>
</Disorder>
```

An individual can be asserted by declaring its identity as different from other individuals.

```
<owl:AllDifferent>
  <owl:distinctMembers
  rdf:parseType="Collection">
    <manufacturer:Disorder rdf:about="#
    Cardiovascular Disorder "/>
    <manufacturer: Disorder rdf:about="#
    Gastrointestinal Disorder "/>
    <manufacturer: Disorder rdf:about="#
    Ophthalmological Disorder "/>
  </owl:distinctMembers>
</owl:AllDifferent>
```

Properties

Properties assert *axioms* (facts) about the members of classes and specific facts about individuals.

The two types of properties are:

- Datatype properties
  - Relate individuals (instances) with data
- Object properties
  - Relate two individuals (instances) together

Note: In OWL DL and OWL Lite, the set of object properties and datatype properties are disjoint.

owl:DatatypeProperty is a proper subclass of the RDF rdf:Property class. Datatype properties link datatypes to individuals.

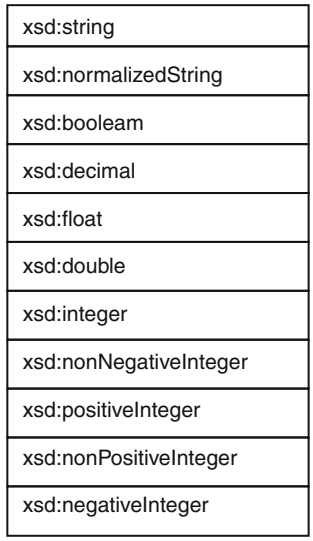


Fig. 3.37 OWL datatypes defined as an XML schema

Datatypes

The OWL datatypes are given in Fig. 3.37. In our previous example of the class "Age" of a person, we linked the nonNegativeInteger datatype to the object property "Years." An instance of the built-in OWLclassowl:ObjectProperty.owl:ObjectProperty is a proper subclass of the RDF rdf:Property class. Object properties relate an individual (instance) to other individuals.

Object properties can also be restrictions of other object properties.

OWL Language Elements

The OWL language semantics are given in Fig. 3.38. Each element is a method of defining or evaluating information in the ontology. For example, OWL:class allows you to define classes of information in the ontology.

OWL Abstract Syntax

In addition to the official OWL exchange syntax, there exists an *OWL abstract syntax*. It is more specific than OWL exchange syntax. The abstract syntax has a less atomic structure that permits easier evaluation and construction of ontological models. It is based on the Extended Backas–Naur Form (EBNF) notation.

Fig. 3.38 OWL semantics allow the definition and classification of information

Class Axioms

The three types of OWL DL class axioms are:

1. General restriction
   - *axiom* ::= '*Class*(' *classID* ['*Deprecated*'] *modality* {*annotation*} {*description*} ')'
   - *modality* ::= '*complete*' | '*partial*'
2. Class assertion
   - *axiom* ::= '*EnumeratedClass*(' *classID* ['*Deprecated*'] {*annotation*} *individualID*}')'
3. Collection of descriptions
   - *axiom* ::= '*DisjointClasses*(' *description description* {*description*} ')'
     | '*EquivalentClasses*(' *description* {*description*}')'
     | '*SubClassOf*(' *description description* ')'

Property Axioms

The property axiom declaration syntax is:

axiom ::= '*DatatypeProperty(*' *datavaluedPropertyID* ['*Deprecated*'] {*annotation*}
{'*super(*' *datavaluedPropertyID* '*)*'}
['*Functional*']
{'*domain(*' *description* '*)*'} {'*range*
(' *dataRange* '*)*'} '*)*'
| '*ObjectProperty(*' *individualvaluedPropertyID* ['*Deprecated*'] {*annotation*}
{'*super(*' *individualvaluedPropertyID* '*)*'}
['*inverseOf(*' *individualvaluedPropertyID* '*)*'] ['*Symmetric*']
['*Functional*' | '*InverseFunctional*' |
'*Functional*' '*InverseFunctional*'|'
Transitive']
{'*domain(*' *description* '*)*'} {'*range*
(' *description* '*)*'} '*)*'
| '*AnnotationProperty(*' *annotationPropertyID* {*annotation*} '*)*'
| '*OntologyProperty(*' *ontologyPropertyID*
{*annotation*} '*)*'

The datatype declaration syntax is:

axiom ::= '*Datatype(*' *datatypeID*
['*Deprecated*'] {*annotation*}*)*'

Restrictions

The restriction axiom declaration syntax is:

restriction ::= 'restriction(' datavaluedPropertyID dataRestrictionComponent {dataRestrictionComponent} ')'
| 'restriction(' individualvaluedPropertyID individualRestrictionComponent {individualRestrictionComponent} ')'
dataRestrictionComponent ::= 'allValuesFrom(' dataRange ')'
| 'someValuesFrom(' dataRange ')'
| 'value(' dataLiteral ')'
| cardinality individualRestrictionComponent ::= 'allValuesFrom(' description ')' | 'someValuesFrom('description')' | 'value('individualID')' | cardinality
cardinality ::= 'minCardinality(' non-negative-integer ')' | 'maxCardinality(' non-negative-integer ')' | 'cardinality (' non-negative-integer ')'

The data range declaration syntax is:

dataRange ::= datatypeID | 'rdfs:Literal' | 'oneOf(' {dataLiteral} ')'

Purpose of Axioms

Axioms can make several properties equivalent or make one property a subproperty of another property. The syntax would look like:

axiom ::= 'EquivalentProperties(' datavaluedPropertyID datavaluedPropertyID {datavaluedPropertyID}')'
| 'SubPropertyOf(' datavaluedPropertyID datavaluedPropertyID ')'
| 'EquivalentProperties(' individualvaluedPropertyID individualvaluedPropertyID {individualvaluedPropertyID} ')'
| 'SubPropertyOf(' individualvaluedPropertyID individualvaluedPropertyID ')'

The W3C maintains a table that maps the OWL abstract syntax against the OWL exchange syntax (too large to include here.)

OWL Language and Its Formal Logic

The semantics here start with the notion of a vocabulary, which can be thought of as the URI references that are of interest in a knowledge base. It is, however, not necessary that a vocabulary consists only of the URI references in a knowledge base.

An OWL vocabulary (V) is a set of URI references, including owl:Thing, owl:Nothing, and rdfs:Literal. Each OWL vocabulary also includes URI references for each of the XML schema nonlist built-in simple datatypes. In the semantics, LV is the (nondisjoint) union of the value spaces of these datatypes.

An abstract OWL interpretation with vocabulary V is a four-tuple of the form: I=<R, S, EC, ER> where

- R is a nonempty set of resources, disjoint from LV
  - S : V → R
  - EC : V → 2^R ∪ 2^LV
  - ER : V → 2^(R × R) ∪ 2^(R × LV)
- S provides meaning for URI references that are used to denote OWL individuals, while EC and ER provide meaning for URI references that are used as OWL classes and OWL properties, respectively.
  - Abstract OWL interpretations have the following conditions having to do with datatypes:

Fig. 3.39 OWL syntax paired with its logical meaning

Syntax - S	EC(S)
owl:Thing	R
owl:Nothing	{ }
rdfs:Literal	LV
complementOf(c)	R - EC(c)
unionOf($c_1 \dots c_n$)	$EC(c_1) \cup \dots \cup EC(c_n)$
intersectionOf($c_1 \dots c_n$)	$EC(c_1) \cap \dots \cap EC(c_n)$
oneOf($i_1 \dots i_n$)	$\{S(i_1), \dots, S(i_n)\}$
oneOf($d_1,l_1 \dots d_n,l_n$)	$\{D(d_1,l_1), \dots, D(d_n,l_n)\}$
restriction(p $x_1 \dots x_n$)	$EC(\text{restriction}(p\ x_1)) \cap \dots \cap EC(\text{restriction}(p\ x_n))$
restriction(p allValuesFrom(r))	$\{x \in R \mid <x,y> \in ER(p) \rightarrow y \in EC(r)\}$
restriction(p someValuesFrom(e))	$\{x \in R \mid \in<x,y> \in ER(p) \rightarrow y \in EC(e)\}$
restriction(p value(i))	$\{x \in R \mid <x,S(i)> \in ER(p)\}$
restriction(p value(d,l))	$\{x \in R \mid <x,D(d,l)> \in ER(p)\}$
restriction(p minCardinality(n))	$\{x \in R \mid \text{card}(\{y : <x,y> \in ER(p)\}) \leq n\}$
restriction(p maxCardinality(n))	$\{x \in R \mid \text{card}(\{y : <x,y> \in ER(p)\}) \geq n\}$

Fig. 3.40 OWL syntax of additional elements with their logical pairings

Syntax - S	EC(S)
restriction(p cardinality(n))	$\{x \in R \mid \text{card}(\{y : <x,y> \in ER(p)\}) = n\}$
Individual(annotation(…) … annotation(…) type(c_1) … type(c_m) pv_1 … pv_n)	$EC(c_1) \cap \dots \cap EC(c_m) \cap EC(pv(pv_1)) \cap \dots \cap EC(pv(pv_n))$
Individual(i annotation(…) … annotation(…) type(c_1) … type(c_m) pv_1 … pv_n)	$\{S(i)\} \cap EC(c) \cap \dots \cap EC(c_m) \cap EC(pv(pv_1)) \cap \dots \cap EC(pv(pv_n))$
pv(p Individual(…))	$\{x \in R \mid \exists y \in EC(\text{Individual}(\dots)) : <x,y> \in ER(p)\}$
pv(p id), for id an individualID	$\{x \in R \mid <x,S(id)> \in ER(p)\}$
pv(p d,l)	$\{x \in R \mid <x,D(d,l)> \in ER(p)\}$

  - If d is the URI reference for an XML schema nonlist built-in simple datatype, then EC(d) is the value space of this datatype.
  - If c is not the URI reference for any XML schema nonlist built-in simple datatype, then EC(c) is a subset of R.
  - If d,l is a datatype,literal pair, then D(d,l) is the data value for l in XML schema datatype d.
- EC is extended to the syntactic constructs of <description>s, <dataRange>s, <individual>s, and <propertyValue>s as follows (see Figs. 3.39 and 3.40):

An abstract OWL interpretation, I, is an interpretation of OWL axioms and facts as given in Figs. 3.41 and 3.42. In Figs. 3.41 and 3.42, optional parts of axioms and facts are given in square brackets ([…]) and have corresponding optional conditions, also given in square brackets.

The effect of an imports construct is to import the contents of another OWL ontology into the current ontology. The imported ontology is the one that can be found by accessing the document at the URI that is the argument of the imports construct. The *imports closure* of an OWL ontology is then the result of adding the contents of imported ontologies into the current ontology. If these contents contain further imports constructs, the process is repeated as necessary. A particular ontology is never imported more than once in this process, so loops can be handled.

Annotations have no effect on the semantics of OWL ontologies in the abstract syntax. An abstract OWL interpretation, I, is an interpretation

Fig. 3.41 OWL axioms and their conditions on interpretation

Directive	Conditions on interpretations
Class(c complete annotation(…) … annotation(…) descr$_1$ … descr$_n$)	EC(c) = EC(descr$_1$) ∩…∩ EC(descr$_n$)
Class(c partial annotation(…) … annotation(…) descr$_1$ … descr$_n$)	EC(c) ⊆EC(descr$_1$) ∩…∩ EC(descr$_n$)
EnumeratedClass(c annotation(…) … annotation(…) i$_1$ … i$_n$)	EC(c) = { S(i$_1$), …, S(i$_n$) }
DisjointClasses(d$_1$ … d$_n$)	EC(d$_i$) ∩ EC(d$_j$) = { } for 1 ≤ i < j ≤ n
EquivalentClasses(d$_1$ … d$_n$)	EC(d$_i$) = EC(d$_j$) for 1 ≤ i < j ≤ n
SubClassOf(d$_1$ d$_2$)	EC(d$_1$) ⊆EC(d$_2$)
DataProperty(p annotation(…) … annotation(…) super(s$_1$) … super(s$_n$) domain(d$_1$) … domain(d$_n$) range(r$_1$) … range(r$_n$) [Functional])	ER(p) ⊆ER(s$_1$) ∩…∩ ER(s$_n$) ∩ EC(d$_1$)×LV ∩…∩ EC(d$_n$)×LV ∩ R×EC(r$_1$) ∩…∩ R×EC(r$_n$) [ER(p) is functional]
IndividualProperty(p annotation(…) … annotation(…) super(s$_1$) … super(s$_n$) domain(d$_1$) … domain(d$_n$) range(r$_1$) … range(r$_n$) [inverse(i)] [Symmetric] [Functional] [InverseFunctional] [Transitive])	ER(p) ⊆ER(s$_1$) ∩…∩ ER(s$_n$) ∩ EC(d$_1$)×R ∩…∩ EC(d$_n$)×R ∩ R×EC(r$_1$) ∩…∩ R×EC(r$_n$) [ER(p) is the inverse of ER(i)] [ER(p) is symmetric] [ER(p) is functional] [ER(p) is inverse functional] [ER(p) is transitive]
EquivalentProperties(p$_1$ … p$_n$)	ER(p$_i$) = ER(p$_j$) for 1 ≤ i < j ≤ n

Fig. 3.42 Additional OWL axioms and their conditions on interpretations

Directive	Conditions on interpretations
SubPropertyOf(p$_1$ p$_2$)	ER(p$_1$) ⊆ER(p$_2$)
SameIndividual(i$_1$ … i$_n$)	S(i$_j$) = S(i$_k$) for 1 ≤ j < k ≤ n
DifferentIndividuals(i$_1$ … i$_n$)	S(i$_j$) ≠ S(i$_k$) for 1 ≤ j < k ≤ n
Individual([i] annotation(…) … annotation(…) type(c$_1$) … type(c$_m$) pv$_1$ … pv$_n$)	EC(Individual([i] type(c$_1$) … type(c$_m$) pv$_1$ … pv$_n$)) is nonempty

of an OWL ontology, O, iff I is an interpretation of each axiom and fact in the imports closure of O. An abstract OWL ontology *entails* an OWL axiom or fact if each interpretation of the ontology is also an interpretation of the axiom or fact. An abstract OWL ontology entails another abstract OWL ontology if each interpretation of the first ontology is also an interpretation of the second ontology. There is no need to create the imports closure of an ontology – any method that correctly determines the entailment relation is allowed.

From the RDF model theory, for V, a set of URI references containing the RDF and RDFS vocabulary, an RDFS interpretation over V is a triple I=< RI, EXTI, SI >. Here, RI is the domain of discourse or universe, i.e., a set that contains the denotations of URI references. EXTI is used to give meaning to properties and is a mapping from RI to sets of pairs over RI×(RI∪LV). Finally, SI is a mapping from V to RI that takes a URI reference to its denotation. CEXTI is then defined as CEXTI(c)={x∈RI |<x,c>∈EXTI(SI(rdf:type))}. RDFS interpretations must meet several conditions, as detailed in the RDFS model theory.

For example, SI(rdfs:subClassOf) must be a transitive relation.

An OWL interpretation, I=< RI, EXTI, SI >, over a vocabulary V, where V includes VRDFS, rdfs:Literal, VOWL, owl:Thing, and owl:Nothing, is an RDFS interpretation over V that satisfies the following conditions (see Fig. 3.43):

Membership in OWL Classes

An OWL construct is a member of an OWL class if it has the following interpretations (see Fig. 3.44). Membership in OWL classes requires the specification of certain types of data.

Fig. 3.43 Examples of OWL interpretations

If E is	then $CEXT_I(S_I(E))=$	With
owl:Thing	IOT	IOT⊆renal insufficiency
owl:Nothing	{}	
rdfs:Literal	LV	
owl:Class	IOC	IOC⊆$CEXT_I(S_I$(rdfs:Class))
owl:Restriction	IOR	IOR⊆IOC
owl:Datatype	IDC	IDC⊆$CEXT_I(S_I$(rdfs:Class))
owl:Property	IOP	intraocular pressure⊆$CEXT_I(S_I$(rdf:Property))
owl:ObjectProperty	IOOP	IOOP⊆IOP
owl:DataTypeProperty	IODP	IODP⊆IOP
rdf:List	IL	IL⊆R_I

IfE is	then$S_I(E)\in$
owl:Thing	IOC
owl:Nothing	IOC
rdfs:Literal	IDC
a datatypeofD	IDC
rdf:nil	IL

Fig. 3.44 OWL datatypes and constructs and their interpretations

Characteristics of the Members of OWL Classes

A description of the defining characteristics of the members of an OWL class is shown in Fig. 3.45.

The next constraints are IFF, which may be harder to deal with in OWL/DL, as they extend the various categories of properties to all of owl:Property (see Fig. 3.46). However, in OWL DL ontologies, you can neither state that an owl:DatatypeProperty is inverse functional nor ask whether it is, so there should be no adverse consequences.

OWL Properties with If and Only If (IFF) Characterizations

There are additional properties with IFF characterizations as shown in Fig. 3.47. These formal constraints help to understand the behavior associated with the application of each of these properties.

OWL Properties with If Characterizations

We will say that l1 is a sequence of y1,…,yn over C iff n = 0 and l1 = SI(rdf:nil) or n > 0 and l1εIL and ∃l2, …, ln ∈ IL such that

<l1,y1 > ∈ EXTI(SI(rdf:first)), y1 ∈ CEXTI(C),
<l1,l2 > ∈ EXTI(SI(rdf:rest)), …,
<ln,yn > ∈ EXTI(SI(rdf:first)), yn ∈ CEXTI(C),
and < ln,SI(rdf:nil) > ∈ EXTI(SI(rdf:rest)).

There are a set of logical relations that define the relationships between members of a class and their properties and constraints (see Fig. 3.48).

Reasoning in OWL

The power of developing a domain ontology in OWL lies in the ability to reason against the ontology. There are many both commercial and public domain reasoners for OWL. An example of one such open source reasoned is the JTP reasoned. The JTP reasoner was developed and maintained at the Knowledge Systems Laboratory (KSL), Computer Science Department, Stanford University by Gleb Frank. It is a Java-based object-oriented reasoned. Its modular design allows for integration with new reasoning models including forward and backward chaining algorithms. Backward chaining reasoners begin with a known goal and derive proofs for the answers returned. Forward chaining reasoners process

Fig. 3.45 Members of OWL classes have these interpretations as do the datatypes, ObjectProperties and DataTypeProperties

IfE is	then if $e \in CEXT_I(S_I(E))$ then
owl:Class	$CEXT_I(e) \subseteq IOT$
owl:Datatype	$CEXT_I(e) \subseteq LV$
owl:ObjectProperty	$EXT_I(e) \subseteq IOT \times IOT$
owl:DatatypeProperty	$EXT_I(e) \subseteq IOT \times LV$

Fig. 3.46 Properties that are true only if they meet certain *if and only if* criteria

If E is	then $c \in CEXT_I(S_I(E))$ iff $c \in IOP$ and
owl:SymmetricProperty	$\langle x,y \rangle \in EXT_I(c) \rightarrow \langle y, x \rangle \in EXT_I(c)$
owl:FunctionalProperty	$\langle x,y_1 \rangle$ and $\langle x,y_2 \rangle \in EXT_I(c) \rightarrow y_1 = y_2$
owl:InverseFunctionalProperty	$\langle x_1,y \rangle \in EXT_I(c) \cap \langle x_2,y \rangle \in EXT_I(c) \rightarrow x_1 = x_2$
owl:TransitiveProperty	$\langle x,y \rangle \in EXT_I(c) \cap \langle y,z \rangle \in EXT_I(c) \rightarrow \langle x,z \rangle \in EXT_I(c)$

Fig. 3.47 Additional properties defined by *if and only if* criteria

If E is	then $\langle x,y \rangle \in EXT_I(S_I(E))$ iff
owl:sameClassAs	$x,y \in IOC \wedge CEXT_I(x) = CEXT_I(y)$
owl:disjointWith	$x,y \in IOC \wedge CEXT_I(x) \cap CEXT_I(y) = \{\}$
owl:samePropertyAs	$x,y \in IOP \wedge EXT_I(x) = EXT_I(y)$
owl:inverseOf	$x,y \in IOOP \wedge \langle u,v \rangle \in EXT_I(x)$ iff $\langle v,u \rangle \in EXT_I(y)$
owl:sameIndividualAs	$x = y$
owl:sameAs	$x = y$
owl:differentFrom	$x \neq y$

Fig. 3.48 For members of a class, this specifies the relationship between the member and its properties and constraints

If there exists	then there exists $y \in IOR$ with
$x \in IOP \wedge w \in IOC \cup IDC$	$\langle y,x \rangle \in EXT_I(S_I(\text{owl:onProperty})) \wedge \langle y,w \rangle \in EXT_I(S_I(\text{owl:allValuesFrom}))$
$x \in IOP \wedge w \in IOC \cup IDC$	$\langle y,x \rangle \in EXT_I(S_I(\text{owl:onProperty})) \wedge \langle y,w \rangle \in EXT_I(S_I(\text{owl:someValuesFrom}))$
$x \in IOP \wedge w \in IOT \cup LV$	$\langle y,x \rangle \in EXT_I(S_I(\text{owl:onProperty})) \wedge \langle y,w \rangle \in EXT_I(S_I(\text{owl:hasValue}))$
$x \in IOP \wedge w \in LV \wedge w$ is a non-negative integer	$\langle y,x \rangle \in EXT_I(S_I(\text{owl:onProperty})) \wedge \langle y,w \rangle \in EXT_I(S_I(\text{owl:minCardinality}))$
$x \in IOP \wedge w \in LV \wedge w$ is a non-negative integer	$\langle y,x \rangle \in EXT_I(S_I(\text{owl:onProperty})) \wedge \langle y,w \rangle \in EXT_I(S_I(\text{owl:maxCardinality}))$
$x \in IOP \wedge w \in LV \wedge w$ is a non-negative integer	$\langle y,x \rangle \in EXT_I(S_I(\text{owl:onProperty})) \wedge \langle y,w \rangle \in EXT_I(S_I(\text{owl:cardinality}))$

assertions and derive conclusions supported by proofs. Dispatchers route assertions and goals to the appropriate reasoning strategy. The JTP reasoned is available for download online at http://www.ksl.stanford.edu/software/JTP/.

Usage models for the JTP reasoner include but are not limited to:

- Embedded reasoner. JTP is embedded in a larger system, preferably implemented in Java to minimize transition expenses.
- Reasoning server. JTP accepts assertions, queries, and control commands over the network from clients. There are no limitations on client implementation.
- Core for a web-based reasoning system. JTP sits on the server; functionality is exported through a web-based interface. This model can be combined with the previous one.
- Core for a client-side reasoning system. JTP sits on the client machine; a human user is accessing it through a UI layer. An extremely spartan implementation of this model is included in standard JTP distribution; it's the class jtp.ui.Console.

SPARQL Query Language

SPARQL, pronounced "sparkle," is an RDF query language; its name is a recursive acronym that stands for ***S****PARQL* ***P****rotocol* ***a****nd* ***R****DF* ***Q****uery* ***L****anguage*. It was standardized by the *RDF Data Access Working Group* (DAWG) of the World Wide Web Consortium and is considered a key Semantic Web Technology. On January 15, 2008, SPARQL became an official W3C Recommendation [45].

SPARQL allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns [46].

Implementations for multiple programming languages exist [45]. "SPARQL will make a huge difference" according to Sir Tim Berners-Lee in a May 2006 interview [47].

SPARQL has four query forms. These query forms use the solutions from pattern matching to form result sets or RDF graphs. The query forms are:

SELECT
Returns all, or a subset of, the variables bound in a query pattern match

CONSTRUCT
Returns an RDF graph constructed by substituting variables in a set of triple templates

ASK
Returns a Boolean indicating whether a query pattern matches or not

DESCRIBE
Returns an RDF graph that describes the resources found

The SPARQL Variable Binding Results XML Format can be used to serialize the result set from a SELECT query or the Boolean result of an ASK query.

Writing a Simple Query

The example below shows a SPARQL query to find the title of a book from the given data graph. The query consists of two parts: the SELECT clause identifies the variables to appear in the query results, and the WHERE clause provides

Data:

```
<http://exampleHealthcareOrganiza-
tion.org/PatientRecords/123456789>
<http://exampleHealthcareOrganiza-
tion.org/PatientName> "John Smith".
```

Query:

```
SELECT ?Patient Name
WHERE
{
<http://exampleHealthcareOrganiza-
tion.org/PatientRecords/123456789>
<http://exampleHealthcareOrgani-
zation.org/PatientName>?Patient
Name .
}
```

the basic graph pattern to match against the data graph. The basic graph pattern in this example consists of a single triple pattern with a single variable (?Patient Name) in the object position.

This query, on the data above, has one solution:

Query Result:

```
Patient Name
"John Smith"
```

SPARQL queries can be used to find semantic information from a triple store of facts or instance data and are a simple yet powerful mechanism for identifying common data from large well-defined warehouses of healthcare data. Semantically encoded healthcare data can be the source of information for our healthcare quality improvement projects. We can and must learn from the practice of medicine to provide every patient with the very best and safest care possible.

Common Logic

Common logic is the ISO standard 24707, for an interchange knowledge representation system. It has the full expressivity of first-order logic; however, it may not be decidable. This means that not all questions that can be asked can be answered. The figure below graphically describes the coverage of the various knowledge representation schemes by the common logic standard. As one can see this forms a Venn diagram with respect to some of the KR systems such as conceptual graphs (see Fig. 3.49).

It is simple to write, you can say:

Names for denoting things

'*JohnSmith*', '*Hospital*','*17*', 'ω'

Predicates for describing the properties of, and relations among, things

Happy(*JohnSmith*), *ReceivesCare*(*Hospital*), $\omega < \omega + 17$

Quantifiers for expressing generality

Hospitals exist – $(\exists x)$*Hospital*(x)

If everyone is happy, Smith is – $(\forall x)$*Happy*(x) $\rightarrow$ *Happy*(*Smith*)

Some infinite ordinal is less than all other infinite ordinals –

$(\exists x)(Ord(x)\ \&\ Inf(x)\ \&\ (\forall y)(Ord(y)\ \&\ Inf(y)\ \&\ x \neq y \rightarrow x < y))$

Features of common logic include:

Strict syntactic typing

Basic lexical elements divided strictly into disjoint classes

Predicate symbols, function symbols, individual constants

Predicates/Fn symbols can take only individual constants as arguments

Individual constants cannot take arguments

Fixed signatures

Each predicate and function symbol takes a fixed number of arguments

Strict semantic typing

Single domain of "individuals"

Individual constants only denote things in the domain

Predicate/function symbols denote things outside the domain

Extensionality

The semantic values of predicate/function symbols are SETS of (n tuples of) individuals

Although common logic is the ISO standard for knowledge representation, this author does

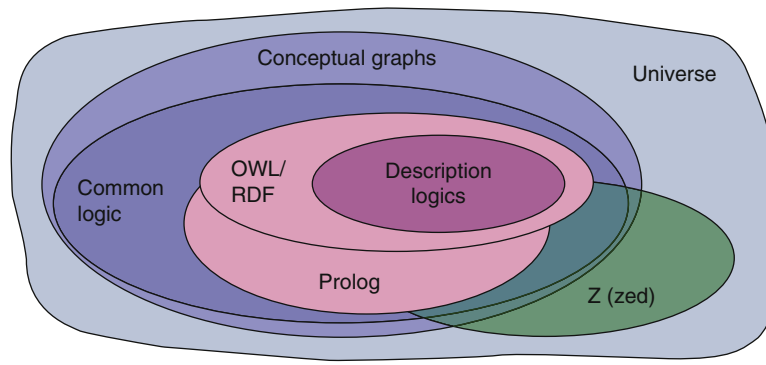


Fig. 3.49 Venn diagram of the coverage of various types of knowledge representation systems for the universe of possible discourse

not know of any common logic reasoners that are available either commercially or in the public domain at the time of this publication. Perhaps common logic reasoners will be more accessible in the future. In the meanwhile, common logic can be used as an interchange format between the different knowledge representation formalisms.

Conclusion

Knowledge representation and its associated logics have the potential to create interoperability between healthcare data within and between healthcare organizations. This information can and should be used to improve our ability to care for our patients. Furthermore we need to employ this approach to speed our ability to learn from our practice and to translate what is learned in research labs to benefit patients more rapidly and efficiently. High-quality healthcare requires a systemized approach to patient care. This is only possible with high-quality, well-defined data to fuel decision support and our research enterprise. With full dissemination of knowledge, we empower our community to continuously learn from our healthcare practices and the practices of our colleagues. This outcome requires rigorous and detailed unambiguous representation of our instance data with well-formed ontologies that will allow local expressivity and yet will aggregate data to their common meaning. Ontology-coded data are the basis for multicenter data sharing, personalized medicine, and improved patient care.

Questions

1. A terminology is just another name for an ontology?
   (a) True
   (b) False
2. An ontology is:
   (a) A formal representation of the knowledge in a domain
   (b) The Web Ontology Language
   (c) A terminology with computable definitions
   (d) a and c
   (e) a, b, and c
3. The UMLS stands for?
   (a) The Uniform Medical Logic System
   (b) The Uniform Medical Language System
   (c) The Unified Medical Language System
   (d) The Unified Medical Logic System
4. All of these are knowledge representation systems except:
   (a) K-Rep
   (b) UMLS
   (c) OWL
   (d) KL-One
5. Part of relations are defined in philosophy as a:
   (a) Ontology
   (b) Topology
   (c) Metrology
   (d) Mereology
6. Description logics are:
   (a) Formal methods for defining concepts in an ontology
   (b) Provide a basis to autoclassify a formal terminology
   (c) Allow the specification of semantic triples
   (d) All of the above
7. All are true of description logics except?
   (a) Subsets of First Order Predicate Logics
   (b) Allow the specification of initial conditions
   (c) Allow the generation of directed acyclic graphs
   (d) All of the above
8. Conceptual graphs are?
   (a) Subsets of First Order Predicate Logic
   (b) Allow the specification of initial conditions
   (c) Allow the generation of directed acyclic graphs
   (d) b and c
   (e) a, b and c
9. In modeling, structural analysis has all these views except?
   (a) Data view
   (b) Modeling view
   (c) Functional view
   (d) Dynamic view
10. In the Unified Modeling Language, protected classes are?
    (a) Only accessible by the class members and members of any subclass

(b) Only accessible by the class members alone
(c) Protected from access by other members of the class
(d) Protected from access by members of their subclasses

11. Foreign keys?
(a) Map information from one language to another
(b) Open security access to tables
(c) Connect two tables within a database by a common data column
(d) Open security access between databases

12. The Object Control Language?
(a) Allows designers to place constraints on operations
(b) Allows designers to assign invariants to objects
(c) Allows designers to assign pre and post conditions to methods
(d) a and b
(e) a, b and c

13. Invariants in OCL are?
(a) Things that are always true about classes but not their subclasses (the transitive closure of subsumption)
(b) Things that are always true about classes and their subclasses (the transitive reflexive closure of subsumption)
(c) Things that are not true for their class but are true for their subclasses
(d) Things that are not true about classes or their subclasses

14. Which statement is true regarding OCL?
(a) A package is a type of model.
(b) A model is made up of multiple packages.
(c) A model is a type of package.
(d) A package cannot contain a model.

15. Which statement(s) are true regarding OWL?
(a) It is a description logic-based knowledge representation language.
(b) It is the language of the Semantic Web.
(c) It is limited to the representation of First Order Predicate Logic.
(d) All of the above.
(e) None of the above.

16. All of the following are species of the OWL language except?
(a) OWL Full
(b) OWL Bright
(c) OWL Description Logic
(d) OWL Lite

17. OWL stands for?
(a) The Ontology Web Language
(b) The Open Web Language
(c) The Web Ontology Learning System
(d) The Web Ontology Language

18. The difference between OWL Full and OWL DL is?
(a) OWL DL does not have the Full set of operations.
(b) OWL DL does not allow same name for objects and properties.
(c) OWL DL has a description logic underpinning the language.
(d) OWL DL is missing the ability to specify invariants.

19. OWL can be represented as all of the following except?
(a) RDF triples
(b) Abstract syntax
(c) XML
(d) Conceptual graph with preconditions

20. The following is/are true about ontologies?
(a) A level one ontology is domain independent.
(b) A level two ontology is domain dependent.
(c) A level three ontology inherits the constraints of the level two and level one ontologies.
(d) A level three ontology is composed of both classes and instances.
(e) All of the above.

References

1. Chomsky N. The logical structure of linguistic theory. New York: Plenum; 1975.
2. Sager N, Lyman M, et al. Natural language processing and the representation of clinical data. J Am Med Inform Assoc. 1994;1(2):142–60.

3. Sager N. Syntactic analysis of natural language. In: Advances in computers, vol. 8. New York: Academic; 1967. p. 153–88.
4. Grishman R, Sager N, Raze C, Bookchin B. The linguistic string parser. In: AFIPS conference proceedings, vol. 42. Montvail: AFIPS; 1973. p. 427–34.
5. Sager N, Gishman R. The restriction language for computer grammars of natural language. Commun ACM. 1975;18:390–400.
6. Wagner JC, Rogers JE, Baud RH, Scherrer JR. Natural language generation of surgical procedures. Int J Med Inform. 1999;53:175–92.
7. Wagner JC, Rogers JE, Baud RH, Scherrer JR. Natural language generation of surgical procedures. Medinfo. 1998;9(Pt 1):591–5.
8. Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux AM, Abrial V, Clavel L, Idir H. Galen: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. Stud Health Technol Inform. 1999;68:901–5.
9. Wroe CJ, Cimino JJ, Rector AL. Integrating existing drug formulation terminologies into an HL7 standard classification using OpenGALEN. Proc AMIA Annu Symp. 2001;766–70.
10. Rector AL. Clinical terminology: why is it so hard? Methods Inf Med. 1999;38(4–5):239–52.
11. Rector AL. The interface between information, terminology, and inference models. Medinfo. 2001; 10(Pt 1):246–50.
12. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004;11(5):392–402.
13. Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. J Am Med Inform Assoc. 2001;8:80–91.
14. Huang Y, Lowe H, Hersh W. A pilot study of contextual UMLS indexing to improve the precision of concept based representation in XML-structured clinical radiology reports. J Am Med Inform Assoc. 2003; 10:580–7.
15. Cooper GF, Miller RA. An experiment comparing lexical and a statistical method for extracting MeSH terms from clinical free text. J Am Med Inform Assoc. 1998;5:62–75.
16. Berrios DC. Automated indexing for full text information retrieval. Proc AMIA Symp. 2000:71–5.
17. Zou Q, Chu WW, Morioka C, Leazer GH, Kangarloo H. IndexFinder: a method of extracting key concepts from clinical texts for indexing. Proc AMIA Symp. 2003:763–7.
18. Srinivasan S, Rindflesch TC, Hole WT, Aronson AR, Mork JG. Finding UMLS metathesaurus concepts in MEDLINE. Proc AMIA Symp. 2002:727–31.
19. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NLM indexing initiative. Proc AMIA Symp. 2000:17–21.
20. Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17–21.
21. Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. J Biomed Inform. 2003;36(4–5):334–41.
22. Humphreys B, McCray A, Cheh M. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc. 1997;4(6):484–500.
23. Campbell KE, Musen MA. Representation of clinical data using SNOMED III and conceptual graphs. Proc Annu Symp Comput Appl Med Care. 1992;16:354–8.
24. Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. J Am Med Inform Assoc Symp. 1998:740–744.
25. Elkin PL, Bailey KR, Chute CG. A randomized controlled trial of automated term composition. J Am Med Inform Assoc Symp. 1998:765–769.
26. Evans DA, Cimino JJ, Hersch WR, Huff SM, Bell D, Group C. Position statement: toward a medical concept representation language. J Am Med Inform Assoc. 1994;1:207–17.
27. Rector A, Rogers J. Ontological issues in using a description logic to represent medical concepts: experience from GALEN. In: IMIA WG6 workshop: terminology and natural language, in medicine, Phoenix; Nov 1999.
28. Elkin PL, Brown SH. Automated enhancement of description logic-defined terminologies to facilitate mapping to ICD9-CM. J Biomed Inform. 2002; 35(5–6):281–8.
29. Rector A. Thesauri and formal classifications: terminologies for people and machines. Methods Inf Med. 1998;37:501–9.
30. Sowa JF. Top-level ontological categories. Int J Hum-Comput Stud. 1996;43:669–85.
31. McDonald FS, Chute CG, Ogren PV, Wahner-Roedler D, Elkin PL. A large-scale evaluation of terminology integration characteristics. JAMIA Suppl. 1999: 864–67.
32. Solomon WD, Roberts A, Rogers JE, Wroe CJ, Rector AL. Having our cake and eating it too: how the GALEN intermediate representation reconciles internal complexity with users' requirements for appropriateness and simplicity, Medical Informatics Group, Department of Computer Science, University of Manchester, Manchester (Tech Report).
33. Rassinoux A-M, Miller R, Baud R, Scherrer J-R. Modeling concepts in medicine for medical language processing. Methods Inf Med. 1998;37(4/5):361–72.
34. Baud RH, Rassinoux A-M, Scherrer J-R. Natural language processing and semantical representation of medical texts. Methods Inf Med. 1992;31:117–21.
35. Baud R, Lovis C, Rassinoux A-M, Scherrer J-R. Alternative ways for knowledge collection, indexing

and robust language retrieval. Methods Inf Med. 1998;37(4/5):315–26.
36. Zanstra PE, Rector AL, Solomon WD, Rush T, Nowlan WA, Bechhofer SK. A terminology server for integrating clinical information systems: the GALEN approach. In: Proceedings of current perspectives in healthcare computing, Harrogate; 1995.
37. Rogers J, Rector A. GALEN's model of parts and wholes: experience and comparisons. In: AMIA annual symposium 2000, Los Angeles; 2000.
38. Elkin PL, Tuttle MS, Keck K, Campbell KE, Atkin GE, Chute CG. The role of compositionality in standardized problem list generation. In: Cesnik B, McCray AT, Scherrer J-R, editors. Ninth world congress on medical informatics (MEDINFO98), Seoul, Korea; 1998. IOS Press, Amsterdam; 1998. p. 660–4.
39. Sowa J. Knowledge representation: logical, philosophical and computational foundations. Pacific Grove: Brooks/Cole; 2000.
40. Carpenter B. The logic of typed feature structure: Cambridge tracts in theoretical computer science. New York: Cambridge University Press; 1992.
41. McGuinness D, Borgida A. Explaining subsumption in description logics. Technical Report LCSR-TR-228, Department of Computer Science, Rutgers University, New Brunswick; 1994. ceur-ws.org/Vol-104/09Liebig-final.pdf.
42. Kabbaj A, Frasson C, Kaltenbach M, Djamen J-Y. A conceptual and contextual object-oriented logic programming: the Prolog++ language. In: Conceptual structures: current practices, Proceedings of the 2nd international conference on conceptual structures, College Park; 1994. Springer, London; 1994.
43. Bagchi A, Wells C. The varieties of mathematical prose – available by anonymous FTP from ftp.cwru.edu. The file mathrite.dvi in the directory math/wells; 1994.
44. Wells C. Communicating mathematics: Useful ideas from computer science. American Mathematical Monthly, 1994;102:397–408.
45. W3C semantic web activity news – SPARQL is a recommendation. W3.org. 2008-01-15. http://www.w3.org/blog/SW/2008/01/15/sparql_is_a_recommendation. Retrieved 2009-10-01.
46. XML and web services in the news. xml.org. 6 October 2006. http://www.xml.org/xml/news/archives/archive.10062006.shtml#5. Retrieved 2007-01-17.
47. Reuters (22 May 2006). Berners-Lee looks for Web's big leap. zdnet.co.uk. Archived from the original on 2007-09-30. http://web.archive.org/web/20070930221904/http://news.zdnet.co.uk/internet/0,1000000097,39270671,00.htm. Retrieved 2007-01-17.

Theoretical Foundations of Terminology

4

Peter L. Elkin

In this chapter, we will discuss the various types of terminologies, their complexity, and the trade-offs associated with the use of each type. Then we will discuss the desirable characteristics of terminologies. This chapter will define the interface between information models and terminological models. The chapter will discuss the power of associating assertional knowledge with terminological knowledge and even the importance of assertional knowledge for the understanding and categorization of terminological knowledge. We will close the chapter with a discussion of the usability of terminologies which is important for anyone who needs to employ terminology in their modeling or characterization of healthcare data.

Types of Terminologies

Terminologies have steadily increased in complexity and rigor since Aristotle developed some of the first biological classification schemes. Initially, observations made of the real world were recorded in one or more contexts. This type of classification was ad hoc and did not intend to connect content from multiple periods of observation. Next, observers and scientists began recording lists of descriptors about a particular topic over a period of observations. One could categorize the wild life on the various islands of the Galapagos or one could record the causes of death in a population. Next, scientists began recording coordinated data in their classifications. For example, scientists could record the length of your shadow at various hours of the day stratified by days of the year while standing at a particular location to learn the angle of incident of light through the day and the seasons.

Terminologies can develop hierarchies of information which is a method to create specializations of information. More complex terminologies can have polyhierarchical representations where more than one hierarchy exists that relates to the same domain. Here, the information from various hierarchies often can be combined to create more complex concepts. For example, a disorder such as a rash may have a location like the bridge of the nose. As these terminologies became useful to more than one person, it became clear that they needed to be definitions. Initially, these were human language definitions that were expanded to become systematized definitions. Here, when complex concepts are formed, they start with a common definition. For example, our definition of a rash of the bridge of the nose would have the same definition as rash specialized with its location.

As terminologies became more complex, it was noted that some concepts should be categorized in, belonged to, multiple hierarchies. For example, colon cancer is both a gastrointestinal disorder and a malignant disorder. As terminologies became larger, it became clear that to know for any given concept all of the places within the

P.L. Elkin, M.D., MACP, FACMI
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_4,

terminology to which it should be categorized was very difficult and specifically to ensure that any new concept is instantiated in all of the correct locations within the terminology.

Formal terminologies (ontologies) were developed to deal with this categorization problem. They provide the mathematical and logical underpinnings necessary to correctly categorize the concepts of a terminology based on their formal definitions. Formal definitions are predicates that related definitional information to a concept. The sum of the predicates related to a concept form the formal computable definition of that concept. In this chapter, we will discuss the types of terminologies and their strengths and weaknesses.

Lists

The simplest form of terminology was a simple list with the name of the terminology being its categorization. This typology is the simplest form of categorization. This listing stemmed from the act of recording one's observations in a domain. For example, if we observed that the human body was composed of a head, chest, an abdomen, arms, legs, a back and buttocks, eyes, ears, a nose, a mouth, hair, and sexual organs, I have created an unorganized list of parts that can be observed looking at the surface of a person. None of the information is ordered, nor is it defined. Where one stops and the other begins is not described, nor is their function.

Lists within lists are the next most common way to describe what one sees in the real world. This is the beginning of specialization and generalization of concepts. Lists within lists become hierarchies as they describe types and subtypes of concepts that are related by a common parent node in the terminology.

Definitions

Systematic Definitions

Systematic definitions are plain language (e.g., English language) definitions of the meaning of concepts. These concepts have a canonical representation of a notion and can have multiple surface forms or synonyms associated with the concept and its unique identifier. These systematic definitions are human readable but not readable by a computer. When done well, they give any human being who wishes to use the concept a clear and unambiguous idea of the meaning of the concept within the terminology. This is important as many statements in the real world have more than one meaning or interpretation; however, for terminological work where consistency is most important, the meaning must remain fixed and clear for all observers.

The definitions are systematic as compound concepts, which are constructed from at least one other concept in the terminology, and have root definitions which do not vary across uses within the terminology. An example of a systematic set of definitions is:

Rash – An eruption of the skin

Rash of the bride of the nose – An eruption of the skin of the bridge of the nose

When writing large sets of definitions, the use of systematic definitions helps to ensure consistency within and across the modelers creating the terminology. Although this is a simple example as terms become complex and may contain more than one primitive term, it can be challenging to keep the definitions systematized and still easy to read. For example, the term "cellulitis of the left foot with osteomyelitis of the third metatarsal without lymphangitis" would draw definitions from each of its five more primitive terms:

1. Cellulitis – An infection of the skin.
2. Left foot – The terminal part of the normal human hindlimb used for ambulation, left side of the body (rather than left lower extremity which is a more complex concept and would not be systematic).
3. Osteomyelitis – Infection of the substance of bone.
4. Third metatarsal bone – A bone in the foot between the tarsal bones and the phalanges, third digit.
5. Lymphangitis – Spread of infection from a source to the lymphatic vessels.

Keeping the language easy to read and consistent (systematic) can be challenging.

Cellulitis of the left foot with osteomyelitis of the third metatarsal without lymphangitis –

> An infection of the skin of the terminal part of the normal human hindlimb used for ambulation, left side of the body that spread to the substance of the bone in the foot between the tarsal bones and the phalanges, third digit without the spread of infection from a source to the lymphatic vessels.

Formal Definitions

Formal definitions are computable. This means that they use a knowledge representation system such as a description logic or a conceptual graph method for representing the definitions of the concepts in the terminology. An example of a description logic representation of a concept would look like:

Acute Myocardial Infarction
  HasMorphology Infarction
  HasFindingSite Myocardium
  HasOnset Acute

Typically, the concept would have a concept code, a canonical representation form, a series of synonyms (0 to many), relationships to other concepts (e.g., Isa relationships or PartOf relations or causal relations; these relations often have a cardinality of 1 to many, such that for example a diagnosis often has more than one possible etiology), and a concept identifier.

This formal definition as described above would be input to a classifier which would check the terminology to know what Isa or PartOf relations were appropriate for this concept and would also check for collisions (other concepts with the same formal definition). Well-formed terminologies should not have collisions.

Hierarchies Within Terminologies

The set of Isa and PartOf relations create the major hierarchies within most terminologies. These hierarchies can be diverse, and concepts can belong (inherit properties from) to multiple hierarchies. This inheritance is the basis for consistency in the knowledge presentation and also a powerful source of distributing information that can be used in your application of the terminology. For example, colon cancer is a malignant disorder and therein inherits all of the properties of the malignant disorders class. Therefore, you do not have to specify what makes something *malignant* as opposed to *benign* for all types of specific malignant disorders. This simplifies greatly the authoring of terminologies and improves the definitional consistency throughout the terminology.

Polyhierarchical terminologies are capable of representing complex relationships and categorizations of knowledge. For example, hierarchies can be used to represent multiple views of the same knowledge that supports its use in different contexts. For example, one could represent disorders by their etiology or by the body system affected.

An example of an Isa hierarchy would be:

Entity
  Disorders
    Cardiovascular disorders
      Ischemic cardiovascular disorders
        Myocardial infarction
          Acute myocardial infarction
            Acute myocardial infarction, anterolateral wall

In this example, each concept would be connected to its parent by the Isa relationship.

Best Practices in Terminology Design and Evaluation

These principals are intended to document the ideas, which are necessary and sufficient to assign value to a controlled health vocabulary. The standard will serve as a guide for governments, funding agencies, terminology developers, terminology integration organizations, and the purchasers and users of controlled health terminology systems toward improved terminological development and recognition of value in a controlled health vocabulary. It is applicable to all areas of health care about which information is kept or utilized. Terminologies should be evaluated within the context of their stated scope and purpose. It is intended to complement and utilize those notions already identified by other national and international standards bodies. This

chapter explicitly refers only to terminologies that are either primarily designed to be used for clinical concept representation or the aspect of a terminology designed to be used for clinical concept representation. This international standard will also provide vocabulary developers and authors with the quality guidelines needed to construct useful, maintainable, controlled health vocabularies. These tenets do not attempt to specify all of the richness, which can be incorporated into a health terminology. However, this chapter does specify the minimal requirements, which if not adhered to will assure that the vocabulary will have limited generalizability and will be very difficult, if not impossible, to maintain. Terminologies, which do not currently meet these criteria, can be in compliance with this standard by putting in place mechanisms to move toward these goals. This standard will provide terminology developers with a sturdy starting point for the development of controlled health vocabularies. This foundation serves as the basis from which vocabulary developers will build robust large-scale reliable and maintainable terminologies.

Terms and Definitions

Terminology

A set of terms representing a system of concepts within a specified domain.

Note: This implies a published purpose and scope from which one can determine the degree to which this representation adequately covers the domain specified.

Controlled Health Vocabulary

A terminology intended for clinical use

Note: This implies enough content and structure to provide a representation capable of encoding comparable data, at a granularity consistent with that generated by the practice within the domain being represented, within the purpose and scope of the terminology.

Classification

A terminology which aggregates data at a prescribed level of abstraction for a particular domain

Note: This fixing of the level of abstraction that can be expressed using the classification system is often created to enhance consistency when the classification is to be applied across a diverse user group, such as is the case with some of the current billing classification schemes.

Ontology

An organization of concepts for which one can make a rational argument.

Note: Colloquially, this term is used to describe a hierarchy constructed for a specific purpose.

Example: A hierarchy of qualifiers would be a qualifier ontology.

Qualifier

A string which when added to a term changes the meaning of the term in a temporal or administrative sense.

Example: “History of” or “recurrent”.

Modifier

A string which when added to a term changes the meaning of the term in the clinical sense.

Example: “Clinical stage” or “severity of illness”.

Canonical Term

A preferred atomic or precoordinated term for a particular medical concept.

Term

A word or words corresponding to one or more concepts.

General Principals

The Basics

Basic characteristics of a terminology influence its utility and appropriateness in clinical applications.

Concept Orientation

The basic unit of a terminology must be a concept, which is the embodiment of some specific meaning, and not a code or character string. Identifiers of a concept must correspond to one and only one meaning, and in a well-ordered vocabulary, only one concept may have that same meaning (ISO/DIS 860). However, multiple terms (linguistic representations) may have the same meaning if they are explicit representations of the same concept. This implies nonredundancy, nonambiguity, nonvagueness, and internal consistency.

Nonredundancy

Terminologies must be internally normalized. There must not be more than one concept identifier in the terminology with the same meaning (ISO 704, E-1284). This does not exclude synonymy; rather, it requires that this be explicitly represented.

Nonambiguity

No concept identifier should have more than one meaning. However, an entry term can point to more than one concept.

Example: MI as myocardial infarction and mitral insufficiency.

Note: Some authors have referred to entry terms as an interface terminology.

Nonvagueness

Concept names must be context free.

Example: "Diabetes mellitus" should not have the child concept "well controlled"; instead, the child concept's name should be "diabetes mellitus, well controlled."

Note: Some authors have referred to context free as context laden.

Internal Consistency

Relationships between concepts should be uniform across parallel domains within the terminology.

Example: If heart valve structures are specified anatomically, the diagnosis related to each structure should also be specified using the same relationships.

Purpose and Scope of a Terminology

Any controlled vocabulary must have its purpose and scope clearly stated in operational terms so that its fitness for particular purposes can be assessed and evaluated. Where appropriate, it may be useful to illustrate the scope by examples or "use cases" as in database models and other specification tools. Criteria such as coverage and comprehensiveness can only be judged relative to the intended use and scope.

Example: A vocabulary might be comprehensive and detailed enough for general practice with respect to cardiovascular signs, symptoms, and disorders but inadequate to a specialist cardiology or cardiothoracic surgery unit. Conversely, a vocabulary sufficiently detailed to cope with cardiology and cardiothoracic surgery might be totally impractical in general practice.

Coverage

Each segment of the healthcare process must have explicit in-depth coverage and not rely on broad leaf node categories that lump specific clinical concepts together. The extent to which the depth of coverage is incomplete must be explicitly specified for each domain (scope) and purpose as indicated in Sect. 4.3 [1].

Example: It is often important to distinguish specific diagnosis from categories presently

labeled "not elsewhere classified" (NEC) or to differentiate disease severity such as indolent prostate cancer from widely metastatic disease.

Comprehensiveness

The extent to which the degree of comprehensiveness is incomplete must be explicitly specified for each domain (scope) and purpose as indicated in Sect. 4.3. Within the scope and purpose, all aspects of the healthcare process must be addressed for all related disciplines, such as physical findings, risk factors, or functional status, across the breadth of medicine, surgery, nursing, and dentistry. This criterion applies because decision support, risk adjustment, outcomes research, and useful guidelines require more than diagnoses and procedures.

Example: Include existing Agency for Healthcare Research and Quality Guidelines and the Health Care Finance Administration (HCFA) mortality model [2].

Mapping

Government and payers mandate the form and classification schema for much clinical data exchange. Thus, comprehensive and detailed representations of patient data within computer-based patient records should be able to be mapped to those classifications, such as ICD-9. This need for multiple granularities is needed for clinical health care as well (ISO/IEC TR 9789). The degree to which the terminology is mappable to other classifications must be explicitly stated [3].

Example: An endocrinologist may specify more detail about a patient's diabetes mellitus than a generalist working in a primary care setting, even though both specialties may be caring for the same patient.

Systematic Definitions

In order for users of the terminology to be certain that the meaning that they assign to concepts is identical to the meaning, which the authors of the vocabulary have assigned, these definitions will need to be explicit and available to the users. Further, as relationships are built into vocabularies, multiple authors will need these definitions to ensure consistency in authorship.

Example: The clinical concept "hypertension" might be defined as a consistently elevated blood pressure and needs to be distinguished from a single "BP > 140/85."

Formal Definitions

A compositional system should contain formal definitions for nonatomic concepts and formal rules for inferring subsumption from the definitions (E-1712).

Explicitness of Relations

The logical definition of subsumption should be defined. The formal behavior of all links/relations/attributes should be explicitly defined. If a looser meaning such as "broader than/narrower than" is used, it should be explicitly stated.

Example: The primary hierarchical relation should be subsumption as exemplified by logical implication: B is a kind of A means all Bs are As.

Reference Terminologies

The set of canonical concepts, their structure, relationships and, if present, their systematic and formal definitions. These features define the core of the controlled health terminology.

Atomic Reference Terminologies

A reference terminology consisting of only atomic concepts and their systematic definitions. In this type of reference terminology, no two or more concepts can be combined to create a composite expression which has the same meaning as

any other single concept contained in the atomic reference terminology.

Colloquial Terminologies

The set of terms, which consist of commonly used entry points, which map to one or more canonical terms within the vocabulary.

Note: These have been called "entry terms" or "interface terminologies" by different authors.

Structure of the Terminology Model

Terminology Structures

Terminology structures determine the ease with which practical and useful interfaces, for term navigation, entry, or retrieval, can be supported (ISO 704, ISO 1087-1, ENV 12264).

Compositional Terminologies

Compositionality

Composite concepts are created from atomic and precoordinated concepts and must be able to be combined to create compositional expressions [4].

Example: "Colon cancer" comprises "malignant neoplasm" and "large bowel" as atomic components. In a compositional system, concept representations can be divided into atomic and composite concept representations.

Composite concept representations can be further divided into "named precoordinated concept representations" and "postcoordinated representation expressions." Within a composite concept, it may be possible to separate the constituents into three categories: "kernel concept," "qualifier (also called "status") concept," and "modifier concept."

Note: A concept is a notion represented by language, which identifies one idea. However, the term "concept" in this technical specification is used to refer to the representation of a concept rather than the thought itself.

Atomic Concept

A representation of a concept that is not composed of other simpler concept representations within a particular terminology. In many cases, atomic concepts will correspond to what philosophers call "natural kinds." Such an entity cannot be meaningfully decomposed. Concepts should be separable into their constituent components, to the extent practical. These should form the root basis of all concepts.

Example: In SNOMED CT, "colon" is a synonym for "large bowel" and "cancer" is a synonym for "neoplasm, malignant." Therefore, the term "colon cancer" is nonatomic as it can be broken down into "large bowel" and "neoplasm, malignant." Each of these two atomic terms has a separate and unique concept identifier, as does the precoordinated term "colon cancer."

Composite Concept

A concept composed as an expression made up of atomic concepts linked by semantic relations (such as roles, attributes, or links).

Precoordinated Concept

Such an entity can be broken into parts without loss of meaning (can be meaningfully decomposed) when the atomic concepts are examined in aggregate. These are representations, which are considered single concepts within the host vocabulary. Ideally, these concepts should have their equivalent composite concepts explicitly defined within the vocabulary (i.e., the vocabulary should be normalized for content).

Example: The term "colon cancer" is nonatomic; however, it has a single unique identifier, which means to the SNOMED-CT that it represents a "single" concept. It has the same status in the vocabulary as the site "large bowel" and the diagnosis "neoplasm, malignant."

Postcoordinated Concept

A composite concept, which is not precoordinated and therefore must be represented as an expression of multiple concepts using the representation language. This is the attempt of a system to construct a set of concepts from within a controlled vocabulary to more completely represent a user's query.

Example: The concept "bacterial effusion, left knee" is not a unique term within the SNOMED-CT terminology. It represents a clinical concept that some patient has an infected left knee joint. As it cannot be represented by a single concept identifier to fully capture the intended meaning, a system would need to build a representation from multiple concept identifiers or lose information to free text.

Types of Atomic and Precoordinated Concepts

Unique concept representations can be classified within a vocabulary into at least three distinct types: kernel concepts, modifiers, and qualifiers (which contain status concepts). This separation allows user interfaces to provide more readable and therefore more useful presentations of composite concepts.

Kernel Concept

This is an atomic or precoordinated concept, which represents one of the one or more main concepts within a precoordinated or postcoordinated composition.

Modifiers and Qualifiers: Terms Which Refine the Meaning of a Kernel Concept

Constituents of a composite concept that refine the meaning of a kernel concept are known as modifiers or qualifiers.

Example 1: "Stage 1a" in the expression "having colon cancer stage 1a" and "brittle, poorly controlled" in the expression "brittle, poorly controlled diabetes mellitus" are examples of qualifiers and modifiers.

In general, these concepts are expressed as a link plus a value ("attribute–value pair"). Terminologies must support a logical structure that can support temporal duration and trend. Attributes must be themselves elements of a terminology and fit into a practical model that extends a terminology.

Example 2: Cancers may be further defined by their stage and histology, have been symptomatic for a specifiable time, and may progress over a given interval.

Attributes are required to capture important data features for structured data entry and pertinent to secondary data uses such as aggregation and retrieval. Kernel concepts can be refined in many ways including a clinical sense, a temporal sense, and by status terms, such as "recurrent."

Normalization of Content

Normalization is the process of supporting and mapping alternative words and shorthand terms for composite concepts. All precoordinated concepts must be mapped to or logically recognizable by all possible equivalent postcoordinated concepts. There should be mechanisms for identifying this synonymy for user-created (new) postcoordinated concepts as well (i.e., when there is no precoordinated concept for this notion in the vocabulary). This functionality is critical to define explicitly equivalent meaning and to accommodate personal, regional, and discipline specific preferences. Additionally, the incorporation of terms as synonyms, represented in a language other than that primarily used in the host vocabulary, can achieve a simple form of multilingual support.

Normalization of Semantics

In compositional systems, there exists the possibility of representing the same concept with multiple potential sets of atoms, which may be linked by different semantic links. In this case, the vocabulary needs to be able to recognize this redundancy/synonymy (depending on your perspective). Therefore, normalization of semantics would recognize all ways that the semantics can be used to represent the same meaning. The extent to which normalization can be performed formally by the system should be clearly indicated.

Example: The concept represented by the term "laparoscopic cholecystectomy" might be represented in the following two dissections:

Surgical Procedure: "Excision"{Has Site Gallbladder}, {Has Method Endoscopic}
and
Surgical Procedure: "Excision"{Has Site Gallbladder}, {Using Device Endoscope}.

Multiple Hierarchies

Concepts should be accessible through all reasonable hierarchical paths (i.e., they must allow multiple semantic parents). A balance between number of parents (as siblings) and number of children in a hierarchy should be maintained. This feature assumes obvious advantages for natural navigation of terms (for retrieval and analysis), as a concept of interest can be found by following intuitive paths (i.e., users should not have to guess where a particular concept was instantiated) [5].

Example: One example of multiple semantic parentage is "stomach cancer" which can be viewed as a "neoplasm" or as a "gastrointestinal disease."

Consistency of View

A concept in multiple hierarchies must be the same concept in each case. The previous example of stomach cancer must not have changes in nuance or structure when arrived at via the cancer hierarchy as opposed to the gastrointestinal disease hierarchy. Inconsistent views could have catastrophic consequences for retrieval and decision support by inadvertently introducing variations in meaning which may be unrecognized and therefore be misleading to users of the system [6].

Explicit Uncertainty

Notions of "probable," "suspected," "history of," or differential possibilities, such as a differential diagnosis list, must be supported. The impact of "certain" versus "very uncertain" information has obvious impact on decision support and other secondary data uses. Similarly, in the case of incomplete syndromes, clinicians should be able to record the partial criteria consistent with the patient's presentation. This criterion is listed separately as many current terminological systems fail to address this adequately.

Representational Form

The representational form of the identifiers within the terminology should be meaningless. Computer coding of concept identifiers must not place arbitrary restrictions on the terminology, such as numbers of digits, attributes, or composite elements. To do so subverts meaning and content of a terminology to the limitations of format, which in turn often results in the assignment of concepts to the wrong location because it might no longer "fit" where it belongs in a hierarchy. These reorganizations confuse people and machines alike, as intelligent navigation agents are led astray for arbitrary reasons. The long, sequential, alphanumeric tags used as concept identifiers in the UMLS project of the National Library of Medicine exemplify well this principle.

Maintenance

Basics of Terminology Maintenance

Technical choices can impact the capacity of a terminology to evolve, change, and remain usable over time.

Context-Free Identifiers

Unique codes attached to concepts must not be tied to hierarchical position or other contexts; their format must not carry meaning. Because health knowledge is being constantly updated, the categorization of health concepts is likely to change. For this reason, the "code" assigned to a concept must not be inextricably bound to a hierarchy position in the terminology so that the code need not change when concepts are hierarchically

reorganized. Changing the "code" may make historical patient data confusing or erroneous.

Example: "Peptic ulcer disease" is now understood as an infectious disease, but this was not always so.

Note: This notion of context-free identifiers is the same as nonsemantic identifiers [7].

Persistence of Identifiers

Codes must not be reused when a concept is obsolete or superseded. Consistency of patient description over time is not possible when concepts change codes; the problem is worse when codes can change meaning. This practice not only disrupts historical analyses of aggregate data but can be dangerous to the management of individual patients whose data might be subsequently misinterpreted.

Note: This encompasses the notion of concept permanence.

Version Control

Updates and modifications must be referable to consistent version identifiers. Usage in patient records should carry this version information. This is true because the interpretation of coded patient data is a function of terminologies that exist at a point in time.

Example: AIDS patients were coded inconsistently before the introduction of the term AIDS.

Terminology representations should specify the state of the terminology system at the time a term is used; version information most easily accomplishes this and may be hidden from ordinary review (ISO 12620, ISO 1087-2, ISO 11179-3, ISO 2382-4) [8, 9].

Editorial Information

New and revised terms, concepts, and synonyms must have their date of entry or effect in the system, along with pointers to their source and/or authority. Previous ways of representing a new entry should be recorded for historical retrieval purposes.

Obsolescence Marking

Superseded entries should be so marked together with their preferred successor. Because data may still exist in historical patient records using obsolete terms, their future interpretation and aggregation are dependent upon that term being carried and cross-referenced to subsequent terms.

Example: Human T-cell leukemia virus type III (HTLV III) to human immunodeficiency virus (HIV)

Recognize Redundancy

Authors of these large-scale vocabularies will need mechanisms to identify redundancy when it occurs. This is essential for the safe evolution of any such vocabulary. This implies normalization of concepts and semantics but specifically addresses the need for vocabulary systems to provide the tools and resources necessary to accomplish this task.

Language Independence

It would be desirable for terminologies to support multilingual presentations. As health care confronts the global economy and multiethnic practice environments, routine terminology maintenance must incorporate multilingual support. While substantially lacking the power and utility of machine translation linguistics, this simplistic addition will enhance understanding and use globally. Have there been translations? What is the expected cost of translation?

Responsiveness

The frequency of updates, or subversions, should be sufficiently short to accommodate new codes and repairs quickly, ideally on the order of weeks.

Evaluation

Basics of Terminological Evaluation

As we seek to understand quality in the controlled vocabularies that are created or used, a standard criteria for the evaluation of these systems is needed. All evaluations must reflect and specifically identify the purpose and scope of the vocabulary being evaluated [10].

Measures of Purpose and Scope

Important dimensions along which purpose and scope should be defined include:

Clinical Area

What is the clinical area of use of the terminology, the disease area of patients addressed, and/or the expected profession of users? Within what parts of health care is the terminology intended to be used and by whom?

Primary Use

What is the primary intended usage of the terminology?

Example: Some areas of usage include reporting for remuneration, management planning, epidemiological research, indexing for bibliographic, web-based retrieval, recording of clinical details for direct patient care, use for decision support, linking of record to decision support, etc.

Persistence and Extent of Use

While some vocabularies are intended, at least initially, primarily for a specific study or a specific site, others are not. If intended to be persistent, what are the means of updating or change management, etc.?

Degree of Automatic Inferencing Intended

Developers should define whether or not and to what degree automatic inferencing is intended. Developers should define whether or not classification is intended to be automatic. Developers should define whether or not it is intended that validation on input be possible and within what limits. Developers should define whether or not postcoordinated expressions are to be accepted and if so, what can be inferred about them and what restrictions must be placed on them (i.e., is formal sanctioning required?).

Transformations (Mappings) to Other Vocabularies

What transformations (mappings) are supported and for what intended purpose? What is the sensitivity and specificity of the transformations?

Example: Transformation for purposes of bibliographic retrieval may require less precision than transformation for clinical usage.

User/Developer Extensibility

Is it intended that the vocabulary be extended by users or application developers? If so, within what limits? If not, what mechanisms are available for meeting new needs as they arise?

Natural Language

Is natural language input or output supported (for analysis or input)? To what level of accuracy?

Other Functions

What other functions are intended?

Example: Linkage to specific decision support systems, linkage to postmarketing surveillance of medications, etc.

Current Status

To what extent is the system intended to be "finished" or work in progress? If different components of the terminology are at different stages of completion, how is this indicated?

Measures of Quality: Terminological Tools

Interconnectivity (Mapping)

Mapping to Vocabulary and Other Coding Systems

To what extent is the vocabulary mappable to other coding systems or reference terminologies?

Vocabulary and Terminological Enhancements

To what extent can the vocabulary accommodate local terminological enhancements?

Vocabulary and Networking

Can the vocabulary server respond to queries sent over a network (LAN, WAN)?

Precision and Recall

What is the precision (positive predictive value) and recall (sensitivity) of information retrieval of clinical content represented with this terminology?

What is the vocabulary's precision and recall for mapping diagnoses, procedures, manifestations, anatomy, organisms, etc. against an established and nationally recognized standard query test set using a standard well-principled method? This should be evaluated only within the intended scope and purpose of the vocabulary system.

Search Engine

Is a standard search engine used in the mapping process?

Usability

Validation of Terminological Usability

Has the usability of the vocabulary been verified?

Interface Considerations

How have interface considerations been separated from vocabulary evaluation?

Prototypes

Has an effective user interface been built? Has the vocabulary been shown to have an effective user interface for its intended use? If not, what are the questions or issues outstanding? Evidence for speed of entry, accuracy, comprehensiveness in practice, etc. with different approaches? If not, is there a proof of concept?

Application Programmer Interfaces (APIs)

Is there support for computer interfaces and system implementers? Is there a demonstrated proof of concept implementation in software? Can it be shown to be usable for the primary purpose indicated? Have there been failed implementations?

Feasibility

If it is intended for use in an electronic patient record (EPR), what are the options for information storage? Has feasibility been demonstrated?

Other Measures of Quality

The generalizability (applicability) of any study design reported (evaluating reported evaluations) should be able to be evaluated.

Healthcare/Clinical Relevance

What is the vocabulary's healthcare/clinical relevance?

Gold Standard

What was the gold standard used in the evaluation?

Study Population

If published population rates are used for comparison, was the study population comparable to the population from which the rates were derived?

Specific Aims of the Study

Were the specific aims clear?

Blinding

Was the study appropriately blinded?

Randomization

Was the test set selection randomized or shown in some sense to be a representative sample of the end-user population?

Test Location Independence

Was the test location different from the developer's location?

Test Location Appropriate for Study Design

How was the test site suited to the study design (tools, resources, etc.)?

Principal Investigator Associations

Was the principal investigator associated with a:
- University
- Academic medical center
- Corporation or company
- Hospital
- Government agency
- Primary care center (health maintenance organization)
- Private practice
- Academic organization

Principal Investigator

Was the principal investigator independent of the vocabulary being evaluated? Does the principal investigator have a track record of publication in this field of study? Have there been any conflicts of interest in performing this research?

Project Completion

Was the project completed in a reasonable period of time?

Sample Size

Power – was the sample size sufficient to show the anticipated effect, should one exist? Statistics – who reviewed the statistical methods?

Personnel

Training Level

What is the average level of training and experience of the study personnel?

Reviewers

Variability – what is the interreviewer variability? Type – what was the type of reviewer (physician, nurse, other clinician, coder, knowledge engineer) used in the study? Independence – were the reviewers blinded to the other reviewers' judgments (i.e., reviewer independence)?

Terminologies and the Application of Assertional Knowledge

Assertional knowledge is the facts related to a domain of study. In health and health care, we have health facts. These facts are important for reasoning over healthcare data and also for categorization of information. For example, the fact that colon cancer commonly starts on adenomatous polyps and can be diagnosed at colonoscopy are both points of assertional knowledge. This knowledge directly influences terminological construction as in ICD9-CM anemia of blood loss is not categorized with other anemias as it is diagnosed at a procedure (which is more expensive to perform than some other tests, e.g., most blood tests) and therefore pays differently as a diagnosis in the hospital than other forms of anemia.

The fact that a flail mitral leaflet can lead to flash pulmonary edema or that 1/3 of all type I diabetics will go on to develop renal insufficiency are examples of assertional knowledge. Facts like these can be used in association with terminologies where the facts are written as axioms and the individual content is coded in the terminology. So, in our example above, the axioms might look like:

(Axiom ::= *Pulmonary Edema* HasEtiology *Flail Mitral Leaflet*) and

(And

Axiom ::= *Type I Diabetes Mellitus* HasComplication *Renal Insufficiency*

Axiom ::= *Renal Insufficiency* HasValue 1/3)

The ability to form axioms holding assertional knowledge that can be viewed or used as rules and to have the elements of the rules codified in the terminology which you are using to represent your clinical or research data provides designers with additional reasoning capability. This capacity for reasoning can be the basis for fully automated electronic quality monitoring [9, 12].

The ability to connect rule-based systems with terminology that can and is used to represent instance data from real clinical patients can be the basis for clinical decision support systems that work to improve the quality and safety of patient care delivery. Generic DL reasoners can take these

axioms and surveil healthcare data to see whether or not the criteria in the rules are met. Feedback can be given in real time to clinicians while they are still with the patient to help them with their diagnoses, workup, and treatment plans.

Usability Evaluation of Terminological Systems Engaged in Recording Clinical Data

Usability studies are essential to the creation of systems that record, represent, and make use of clinical data. Without formal testing, too often systems alter the meaning of data as it is recorded, and this can lead to errors that effect not only research results but also clinical outcomes.

A usability study evaluates how a particular process or product works for individuals (see Fig. 4.1) [13]. Optimally, one would test a population of individuals who are a sample of typical users of the type of process or product being tested. It should be stated clearly to participants that the purpose of the study is to evaluate the process or product and not the individual participant [14]. Usability sessions can be videotaped from multiple angles (including the computer's screen image), and participants are encouraged to share their thoughts orally as they progress through the scenarios provided ("think aloud") [15]. This helps to define the participants' behavior in terms of both their intentions and their actions [16]. For example, in our study, we had the user identify what information they were looking for before they initiated their search. We will monitor what is entered into the program, and we are able to view the information retrieved. Then we record the degree to which the clinician–user felt that they were satisfied with the information that they had obtained [17].

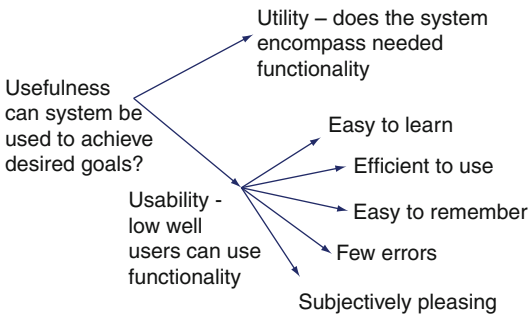


Fig. 4.1 Attributes of usefulness, as exemplified by bench testing. Here, we also depict the axes of usability. These depictions serve to emphasize the goals and challenges to the design of a well-formed web (hypertext) environment

To accomplish a valid study, one must follow a specific protocol and have multiple participants (typically 6–12) interact with the system using the same set of scenarios [18]. It is important that the design team be able to observe multiple participants if they are to become informed by the study. The scenarios should reflect the way the system being tested is actually going to be utilized [19]. The closer the study design can mimic the true end-user environment, the more validity the results of the study will have [20]. In this manner, developers ascertain characteristics of their web environment that are functional, need improvement, fit user expectations, miss expectations, fail to function, or are opportunities for development [21].

Interoperability

Interoperability is a common understanding of the meaning of data between a sending and receiving computer system [22]. The level of interoperability required varies with application needs. The specification of data in enough detail to create a common shared meaning between organizations is a complex task as systems work within organizational and human factors contexts as well as having specific technical requirements. *Aequus communis sentential*, the title of a manuscript published by Elkin et al., translates from Latin to the "level of common meaning." In a manuscript, we have defined and validated an ontology of interoperability [23]. The scale asks reviewers of a specification to define its level in terms of syntactic, semantic, and pragmatic interoperability. We tested the scale by having five medical informaticians rate a set of ANSI standard specifications, and we report the interrater variability of the interoperability rating scheme. Our interoperability rating ontology has high interrater reliability and

is a relatively simple mechanism for comparing the levels of interoperability afforded by different specifications or the same specification over multiple versions.

Ratings using this scale will help consumers of health informatics standard to better understand the level of interoperability provided by any particular specification. Further, we believe that the use of this scale will help these same consumers, who are faced with the choice of which standards to implement, to compare the relative levels of syntactic, semantic, and pragmatic interoperability provided by each of the specifications under review.

An example of the scale's usage is that the ASTM Continuity of Care Record (CCR) was judged by the reviewers of our study most commonly to have a 6e ß level of interoperability. This helps individuals trying to choose a standard to use to compare the interoperability provided by each standard that is under consideration. Standard developers can use this scale to plan for higher levels of interoperability in their next release.

Interoperability scale	
Syntactic interoperability	
a	Headings (e.g., section of the clinical record)
b	Select fields are delimited
c	B plus data types are fixed and reliable
d	C plus numbers are broken out along with values (e.g., blood pressure and values are diastolic and systolic values)
e	D plus hierarchical structure of data without nonhierarchical relationships between fields (e.g., XML structures)
f	E plus nonhierarchical relationships can be specified
Semantic interoperability	
1	Free text
2	Free text with fixed data types
3	Codification of data by local codes
4	Codification of data by nationally standard aggregate codes
5	Codification of data by nationally standard detailed coding system allowing both atomic and precoordinated concepts
6	Codification of data by nationally standard detailed coding system allowing postcoordination (based on formal logic)
7	Model-based knowledge representation with local codes
8	Model-based knowledge representation with nationally standard aggregate codes
9	Model-based knowledge representation with nationally standard detailed coding system allowing both atomic and precoordinated concepts
10	Model-based knowledge representation coordinated semantically nationally standard detailed coding system allowing postcoordination (based on formal first order logic)
11	Model-based knowledge representation coordinated semantically nationally standard detailed coding system allowing postcoordination with support for context (based on formal higher-order logic)
Pragmatic interoperability	
œ	Currently available and easily implemented
ß	Currently available but with barriers to implementation
¥	Barriers could be overcome within one year
∂	Barriers could be overcome within three years
€	Barriers could be overcome within ten years
£	Would take longer than ten years to achieve
µ	Not practically achievable
∞	Not possibly achievable

Interoperability is essential for information about patients to be shipped from one computer to another in a reliable and computable manner. The information once transferred should be adequate to drive the local clinical decision support software, thereby helping to improve patient safety and optimize patient outcomes. We have presented a scale with good interrater agreement which can help implementers of healthcare standards to better understand the level of interoperability provided by standard specifications that they are considering implementing. This transparency, we believe, will help mitigate the risk of choosing a healthcare standard and in that regard will fuel adoption of standards in health IT solutions.

Questions

1. The simplest form of hierarchies are?
   (a) Ordinal lists
   (b) Lists within lists
   (c) Groups of lists
   (d) Nonordinal lists
2. A canonical term in a terminology is?
   (a) The first term in the terminology
   (b) The last term in the terminology
   (c) The preferred term for a concept
   (d) A synonym of the preferred term for a concept
3. Concept orientation of a terminology means?
   (a) That the concept is the basic organization of meaning
   (b) The concept is an abstract notion
   (c) The concept orientation is a hierarchy within the terminology
   (d) The concept identifier is the meaning of the concept
4. Vagueness in a terminology occurs when?
   (a) A term is not well written
   (b) A concept is not well written
   (c) There is more than one concept with the same meaning
   (d) The concept is context dependent
   (e) There is more than one meaning for a concept
5. Redundancy happens when?
   (a) A term is not well written
   (b) A concept is not well written
   (c) There is more than one concept with the same meaning
   (d) The concept is context free
   (e) There is more than one meaning for a concept
6. Ambiguity happens in a terminology when?
   (a) A term is not well written
   (b) A concept is not well written
   (c) There is more than one concept with the same meaning
   (d) The concept is context free
   (e) There is more than one meaning for a concept
7. Internal consistency within a terminology means that?
   (a) Relationships between concepts should be uniform across parallel domains within the terminology
   (b) Relationships follow the Isa relation
   (c) Relationships follow the PartOf relation
   (d) Relationships between concepts should be uniform within parallel domains across the terminology
8. The coverage of the terminology = ?
   (a) The number of concepts in the terminology that cover concepts from the real world/the number of concepts in the real world
   (b) The number of domain concepts in the terminology that cover concepts from the real world/the number of concepts in the real world
   (c) The number of domain concepts in the terminology that cover concepts from the same domain from the real world/the number of concepts in the domain within the real world
   (d) The number of concepts in the domain of the real world/the number of concepts in the terminology
9. Mapping between terminologies is accomplished by?
   (a) Mapping the meaning of concepts of the two terminologies
   (b) Mapping the terms of the two terminologies
   (c) Mapping the hierarchies of the two terminologies
   (d) Mapping the relationships of the two terminologies

10. A composite concept is?
    (a) Any precoordinated concept
    (b) Any postcoordinated Concept
    (c) Any concept with a formal definition
    (d) Any pre- or postcoordinated concept
11. Atomic concepts are?
    (a) Concepts that are made up of atoms
    (b) Concepts that cannot be further decomposed in the terminology
    (c) Concepts that have been used in precoordinated concepts
    (d) Concepts that have been used in postcoordinated concepts
12. Concept is to term as?
    (a) Watermelon is to seed
    (b) House is to furniture
    (c) Idea is to name
    (d) Thought is to being
13. A synonym is an?
    (a) Abbreviation
    (b) Acronym
    (c) Homonym
    (d) a and b
    (e) a, b, and c
14. Precoordinated concepts are?
    (a) Those concepts that are created before the terminology is created
    (b) Those concepts that can be defined by more than one concept in the terminology
    (c) Those concepts that are created before the terminology is coordinated
    (d) Those concepts that are created before the terminology is finalized
15. Postcoordinated concepts are?
    (a) Those that are created from multiple concepts in the terminology and that do not themselves exist as concepts within the terminology
    (b) Those that are created from multiple concepts in the terminology that do themselves exist as concepts within the terminology
    (c) Those that are created from multiple concepts in the terminology and joined with concepts outside the terminology
    (d) Those that are created from multiple concepts in the terminology and then mapped to other terminologies
16. A kernel concept is?
    (a) A concept that is at the center of the terminology
    (b) A concept that represents the main meaning of a precoordinated concept
    (c) A concept that represents the main meaning of a pre- or postcoordinated concept
    (d) A concept that represents the main meaning of a postcoordinated concept
17. A modifier is?
    (a) A concept that modifies the meaning of coordinated concept in a clinical sense
    (b) A concept that modifies the meaning of precoordinated concept in a clinical sense
    (c) A concept that modifies the meaning of postcoordinated concept in a clinical sense
    (d) A concept that modifies the meaning of an atomic concept in a clinical sense
18. A qualifier is?
    (a) A concept that modifies the meaning of coordinated concept in a temporal or administrative sense
    (b) A concept that modifies the meaning of precoordinated concept in a temporal or administrative sense
    (c) A concept that modifies the meaning of postcoordinated concept in a temporal or administrative sense
    (d) A concept that modifies the meaning of an atomic concept in a temporal or administrative sense
19. Consistency of view means that?
    (a) The interface to the terminology should not change between terminologies
    (b) The direction of the hierarchies should not change within a terminology
    (c) The meaning of a terminology should not change over time
    (d) The concepts should have the same descendants regardless of their parentage
20. Explicitness of relations means that?
    (a) Relations should have the same meaning throughout the terminology
    (b) Relations should have the same meaning and be used consistently throughout the terminology

(c) Relations should be used consistently throughout the terminology
(d) Relations should only change their meaning under specific circumstances within a terminology

21. A rule that governs the sign of concepts is?
(a) Concept orientation
(b) Nonredundancy
(c) Explicit uncertainty
(d) Nonvagueness
(e) Nonambiguity

22. Normalization of content is?
(a) The process of supporting and mapping alternative words and shorthand terms for composite concepts
(b) The process of mapping concepts from one terminology to another terminology
(c) The process of mapping the semantics of one terminology to another terminology
(d) The process or recognizing all ways that the semantics can be used to represent the same meaning

23. Semantic normalization is?
(a) The process of supporting and mapping alternative words and shorthand terms for composite concepts
(b) The process of mapping concepts from one terminology to another terminology
(c) The process of mapping the semantics of one terminology to another terminology
(d) The process or recognizing all ways that the semantics can be used to represent the same meaning

24. Context-free identifiers are?
(a) An unbinding of the concept and the identifier
(b) Concept identifiers whose format does not carry meaning
(c) An unbinding of the concept from the terminology
(d) Concept identifiers that do not associate with concepts that carry meaning

25. Persistence of identifiers means?
(a) That identifiers are only reused when the concept is deleted
(b) That identifiers are deleted when the concept is deleted
(c) That the identifiers are never deleted
(d) That the identifiers are always reused

26. Obsolescence marking is exemplified by?
(a) Marking a concept that is old
(b) Marking a concept that is deleted
(c) Marking a concept whose surface form has changed its meaning
(d) Marking a concept that has been deleted or whose surface form has changed its meaning

27. Obsolescence marking requires?
(a) Just the marking of the concept
(b) Marking the concept and pointing to the new concept with the modified meaning
(c) Marking the concept and showing its age
(d) Marking its age and showing all the concepts with that age

28. Language independence for a terminology means?
(a) That it supports multiple languages
(b) That it does not need a language to represent its knowledge
(c) That it has a concept identifier
(d) That it mixes multiple languages as synonyms

29. Precision is?
(a) The true positive over the true positive plus the false positive rates
(b) The true positive over the true positive plus the false negative rates
(c) The true negative over the true positive plus the false positive rates
(d) The true negative over the true positive plus the false negative rates

30. Recall is?
(a) The true positive over the true positive plus the false positive rates
(b) The true positive over the true positive plus the false negative rates
(c) The true negative over the true positive plus the false positive rates
(d) The true negative over the true positive plus the false negative rates

31. Assertional knowledge is?
(a) Ontological knowledge
(b) Facts about the terminology
(c) Facts about the domain expressed in the terminology
(d) Axioms defining the terminology

32. Usability testing is appropriate?
    (a) When you can find typical participants
    (b) When you are willing to make changes to the system
    (c) When you run at least six participants through a set of structured scenarios
    (d) All of the above
33. Usability testing should be performed?
    (a) As soon as you think of a system design
    (b) As soon as you have a design specification to test
    (c) As soon as you have a version of the system ready for beta testing
    (d) As you need information to make design choices throughout the development lifecycle
34. Steps in running a usability experiment include all except?
    (a) Cognitive task analysis
    (b) Specifying a new system design based on the usability data
    (c) Developing typical scenarios for the study
    (d) Recruiting typical participants for the study
35. All of these are potentially useful usability results except?
    (a) Making leadership aware of the good work performed by the development team
    (b) Making the development team aware of usability errors
    (c) Making the development team aware of design elements that worked well
    (d) Making the usability team and the design team aware of issues for further usability studies
36. Which are scenarios where you should not perform usability testing?
    (a) When the study cannot be accommodated within the lab space
    (b) When the study results will not be used to improve the system or make a purchasing decision
    (c) When typical participants cannot be recruited for the study
    (d) All of the above
37. All of the following are linkages between patient safety and usability of clinical systems except?
    (a) Confusing labels on the screen
    (b) Inability to link together relevant clinical data
    (c) Usability errors identified in a usability study
    (d) Unpleasant looking screens (graphical user interfaces)
38. The human-centered design development lifecycle is?
    (a) ISO 12207
    (b) ISO 13407
    (c) ISO 17117
    (d) ISO 9000
39. Usability of systems is composed of all except?
    (a) Ease of error correction
    (b) Efficient to use
    (c) Subjectively pleasing
    (d) Few errors
40. Types of human factors engineering include all except?
    (a) Low fidelity prototyping
    (b) Contextual inquiry
    (c) Expert evaluation
    (d) Competitive usability evaluation
41. Usability engineering is?
    (a) A usability laboratory
    (b) A process
    (c) A randomized controlled trial
    (d) An observational study

References

1. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4/5):394–403.
2. Cote RA, Rothwell DJ. The classification-nomenclature issues in medicine: a return to natural language. Med Informatics. 1989;14(1):25–41.
3. Rocha RA, Rocha BH, Huff SM. Automated translation between medical vocabularies using a frame-based Interlingua. In: Proceedings of the annual symposium of computer applications in medical care, Washington, D.C.;1993. p. 690–4.
4. Bernauer J, Franz M, Schoop D, Schoop M, Pretschner DP. The compositional approach for representing medical concept systems. Medinfo. 1995;8(Pt 1):70–4.
5. Campbell KE, Musen MA. Representation of clinical data using SNOMED III and conceptual graphs. In: Proceedings of the annual symposium on computer applications in medical care, Baltimore; 1992. p. 354–8.
6. Rossi Mori A, Galeazzi E, Gangemi A, Pisanelli DM, Thornton AM. Semantic standards for the representation of medical records. Med Decis Mak. 1991; 4(Suppl):S76–80.

7. Tuttle MS, Olson NE, Campbell KE, Sherertz DD, Nelson SJ, Cole WG. Formal properties of the metathesaurus. In: Proceedings of the annual symposium on computer applications in medical care, Los Alamitos;1994. p. 145–9.
8. Campbell KE, Cohn SP, Chute CG, Rennels G, Shortliffe EH. Galapagos: computer-based support for evolution of a convergent medical terminology. JAMIA. 1996;SympSuppl:269–73.
9. Cimino JJ. Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. Methods Inf Med. 1996;35(3):211–7.
10. Elkin PL, Chute CG. ANSI-HISB code set evaluation criterion survey; 1998, Minutes ANSI-HISB meeting 4/98.
11. Brown SH, Elkin PL, Rosenbloom ST, Fielstein E, Speroff T. eQuality for all: Extending automated quality measurement of free text clinical narratives. AMIA Annu Symp Proc. 2008;6:71–5.
12. Brown SH, Speroff T, Fielstein EM, Bauer BA, Wahner-Roedler DL, Greevy R, Elkin PL. eQuality: automatic assessment from narrative clinical reports. Mayo Clin Proc. 2006;81(11):1472–81.
13. Nielsen J. Usability engineering. New York: Academic Press; 1993.
14. Preece J, Rogers Y, Sharp H, Benyon D, Holland S, Carey T. Human-computer interaction. New York: Addison-Wesley Publishing Company; 1994.
15. Hix D, Hartson HR. Developing user interfaces: ensuring usability through product and process. New York: Wiley; 1993.
16. Coble JM, Karat J, Orland MJ, Kahn MG. Iterative usability testing: ensuring a usable clinical workstation. In: Masys DR, editor. Proceedings of the 1997 AMIA annual fall symposium. Philadelphia: Hanley & Belfus Inc.; 1997. p. 744–8.
17. Kushniruk AW, Patel VL, Cimino JJ. Usability testing in medical informatics: cognitive approaches to evaluation of information systems and user interfaces. Proc AMIA Annu Fall Symp. 1997:218–22.
18. Nielsen J. Estimating the number of subjects needed for a thinking aloud test. Int J Hum Comput Stud. 1994;41:385–97.
19. Patel VL, Ramoni MF. Cognitive models of directional inference in expert medical reasoning. In: Feltovich PJ, Ford KM, Hoffman RR, editors. Expertise in context: human and machine. Cambridge: MIT Press; 1997. p. 67–99.
20. Weir C, Lincoln MJ, Green J. Usability testing as evaluation: development of a tool. In: Cimino JJ (ed), Proceedings of the twentieth annual symposium of computer applications in medical care, Washington, D.C.;1996. p. 870.
21. Kushniruk A, Patel V, Cimino JJ, Barrows R. Cognitive evaluation of the user interface and vocabulary of an outpatient information system. In: Cimino JJ (ed), Proceedings of the twentieth annual symposium of computer applications in medical care, Washington, D.C.;1996. p. 22–26.
22. Brailer DJ Interoperability: the key to the future health care system. Health Aff (Millwood) Suppl Web Exclusives, 2005:W5-19–21.
23. Elkin PL, Froehling D, Bauer BA, Wahner-Roedler DL, Rosenbloom ST, Bailey K, et al. Aequus communis sententia: defining levels of interoperability. Medinfo. 2007;12(Pt 1):725–9.

Compositionality: An Implementation Guide

5

Peter L. Elkin and Steven H. Brown

Introduction

Vocabulary construction and organization is an essential part of a functional Electronic Health Record [1]. Concept level understanding of our day-to-day clinical practice will enable more accurate and more available outcomes research, evidence-based medicine, and effective cost management of medicine without a decline in service. This promise is hampered by the lack of a robust clinically relevant large-scale vocabulary, with a structure, which supports synonymy, multiple ontologies, semantic relationships, and compositionality [2, 3]. In recent years, many accomplishments have been made in the areas of synonymy, ontology, and semantic relationships. Compositionality is an area in which the underlying theory and practical implementation are relatively less well developed despite generally acknowledged payoffs for accurate data representations. In this chapter, we will review the promise and many challenges of compositionality. Toward that end, we will define compositionality and what constitutes a compositional system. We will define a set of rules for generating safe compositional expression (Desiderata for Composition). We will define methods for using multiple terminologies in a composite compositional expression, and we will present formalisms for defining the logical underpinnings that makes the use of compositional systems safe and scalable.

We will define a method for safely using compositional expressions in your institution, and we will provide examples of compositional expressions. We will provide a method for determining which content and semantics from two terminologies can be safely used together to form composite compositional expressions. At the end of this chapter, the reader will understand the need for compositional systems in health care.

This chapter should inform the interested reader with regard to the content and semantics needed for the safe and effective use of composition toward an expressive and accurate method for data representation for health care.

But first we must ask:

What Is Compositionality?

Compositions are expressions made up of sets of concepts joined by relationships, usually in a tree structure defined using a description logic; however, other mechanisms such as directed acyclic graphs (e.g., Conceptual Graphs) are also acceptable logical mechanisms for representing the same data. Each concept must be joined in an appropriate way via a "relation" with dependent concepts. We define the concept to be specified as the oper-

P.L. Elkin, M.D., MACP, FACMI (✉)
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

S.H. Brown, M.D., MS, FACMI
Department of Veterans Affairs and Vanderbilt, Department of Biomedical Informatics, Nashville, TN, USA

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_5,

Table 5.1 Links NDF-RT concepts and relations to SNOMED CT concepts

Concept	New linking relation	Concept
Drugs	Treatment_for	Disease or finding
Drugs	Prevent	Disease or finding
Drugs	Cause_of	Disease or finding
Physiologic effects	Sign_of	Disease
Physiologic effects	Treatment_for	Disease
Mechanism of action	Treatment_for	Disease or finding
Mechanism of action	Cause_of	Disease or finding
Mechanism of action	Prevent	Disease or finding

and and the concept that increases the granularity of the operand as the specification (see Table 5.1).

Example: cellulitis of the foot	
Cellulitis	Operand – Concept
Has_Finding_Site	Relation (or Role)
Foot	Specification – Concept

Definitions

Atomic Concept: A notion represented by language, which identifies one idea. Such an entity cannot be broken into parts without the loss of meaning.

Example: In the UMLS Metathesaurus, colon is a synonym for large bowel, and cancer is a synonym for neoplasm, malignant. Whereas colon cancer is nonatomic as it can be broken down into "large bowel" and "neoplasm, malignant." Each of these two more atomic terms has a separate and unique Concept Unique Identifier (CUI).

Precoordinated Concept: A notion represented by language, which identifies one idea. Such an entity can be broken into parts without loss of meaning when the atomic concepts are examined in combination. These are terms which are considered single concepts within the host vocabulary.

Example: Colon cancer is nonatomic; however, it has a single CUI, which means to the Metathesaurus that it represents a "single" concept. It has the same status in the vocabulary as the site "large bowel" and the diagnosis "neoplasm, malignant."

Postcoordinated Concepts: A notion represented by language and a set of codes (concept level identifiers), which identifies one idea. This is the attempt of a system to construct a set of concepts from within a controlled vocabulary to more completely represent a user's query.

Example: The concept "Status-Post CABG" is not a unique term within the UMLS Metathesaurus. It represents a clinical concept that some patient has already had heart surgery. As it cannot be represented by a single CUI, to fully capture the intended meaning, a system would need to build a representation from multiple CUIs or lose information to free text.

User-Directed Coordination of Concepts: A notion represented by language and a set of codes (concept level identifiers), which identifies one idea. The user chooses this set of concepts, usually via a Graphical User Interface, and usually, we envision that this would occur at the point-of-care. This is the attempt of a user to represent a clinical concept using a set of concepts, whether they are atomic, precoordinated, or postcoordinated concepts. The user or clinician's focus is to most fully capture the meaning that they wish to record regarding their patients.

Example: A GUI, which enables users to combine concepts in a meaningful way. This in our view implies a robust representational schema. Such a schema would facilitate an understanding of these compound structures and their relative locality within the canonical vocabulary. These structures should be nonredundant and should facilitate vocabulary maintenance.

Normalization of Content and Semantics: Normalization is defined as the ability to identify every representational format that confers the same meaning as being equivalent (i.e., unambiguous representation).

Why Does Compositionality Matter?

Users demand the ability to form problem statements that represent the concepts of their practice. We do not and cannot anticipate everything

a clinician might wish to say about a patient. Thus, without fully functional natural language processing, we cannot represent clinical medicine completely within a well-formed controlled vocabulary. One solution is compositionality. The expressive power of compositionality demonstrated in the literature [19] stems from the observation that all of the complex and varied statements that clinicians make regarding their patients can be derived from a manageable number of atomic concepts (estimated to fall somewhere between 20,000 and 1,000,000) [4–6].

Implementing Compositionality: An Overview

Once a decision has been made to permit compositional expressions, an approach to implementation must be developed. This approach should have several components. A philosophical issue that must be considered early is deciding to what extent compositional terms will be precoordinated, versus just-in-time postcoordination. Other components of an implementation approach include development of syntactic and semantic methods to create compositional expressions, scalable methods of delivering software that supports the creation and management of compositional expressions, and an approach to using existing terminologies as components of compositional expressions.

A stepwise approach to this goal includes the following desiderata:

1. *Semantic Independence*: Only use semantics that are independent (nonoverlapping in meaning).
2. *Uniformity of Semantics*: That within the terminology the semantics are instantiated everywhere that they apply.
3. *Logical Consistency*: That the formalism for creating compositional expressions is logically rigorous and is applied in a consistent way based on a formal set of rules (see section "Using OWL as a formal language to represent knowledge unambiguously").
4. *Semantic Normalization*: That there is a process for normalizing the terminology (identifying all of the different precoordinated or postcoordinated compositions that have the same meaning) (see section "Normalization of both content and semantics").
5. *Computational Normalization*: That there is a classifier associated with the formal terminology so that new terms can find their appropriate place in the ontology (see section "Formal knowledge representation").
6. *Colloquial Normalization*: All grammars used in conjunction with the compositional terminology are capable of preventing the generation of ambiguous expressions (see section "Definition").

Associated with these basic principles of composition, we believe there is the need for software that is capable of automatic or user-directed generation of unambiguous compositional expressions for both information capture and retrieval. Automatic generation means the ability to build compositions from free-text input alone. User-directed compositions rely on more usual structured data entry techniques (see section "Architectures to implement compositionality").

Precoordination Versus Postcoordination Spectrum

Once a decision has been made to permit compositional expressions, terminology designers and implementers must subsequently address the issues associated with precoordination vs. postcoordination. As defined in section "Definitions," precoordinated compositions are present in the distributed version of the terminology and available to all users. Postcoordinated terms are not present in the distributed terminology and are created on an ad hoc basis by end users. The decision to permit ad hoc generation of compositions should first be considered. An alternative is to allow users to request novel compositions from a centralized terminology "authority" and to make them wait for the novel composition until the distributed terminology is updated (this is slower but will limit the number of expressions created). If ad hoc postcoordinations are permitted, systems designers may wish to consider methods to review the newly created compositions and consider them for inclusion in subsequent versions of the distributed terminology. The extent to which pre-

coordinated terms are to be supported in the distributed terminology also merits consideration. Clearly, some precoordinated compositions are essential (e.g., "colon cancer"). On the contrary, precoordinating "history of" with every possible problem that could occur historically is wasteful and contributes to increased maintenance overhead.

Compositional systems facilitate data representation using both precoordinated and postcoordinated expressions. This greatly increases the expressivity of the terminology. However, it makes normalization of both the content and semantics essential if one is to avoid creating concepts that are represented differently but have the same meaning (unrecognized ambiguity). One solution would be to separate the truly atomic terms and their ontology from the compositions and their relationships. This multiaxial schema for vocabulary design is clearly controversial.

An example of this type of construction would be "coronary artery disease (CAD) Status Post CABG." Here we have multiple atomic concepts. On first cut, the coronary artery disease can be separated from the s/p CABG. This is only possible if there exists a mechanism for reconstruction. This is clinically very important because the patient with CAD s/p CABG is clearly a different presentation than a patient with CAD without a history of prior cardiac surgery. More controversial is the corollary that the construction of coronary artery and atherosclerotic vascular disease should be an equivalent concept to CAD.

Clearly, we would not want s/p CABG precoordinated in a reference terminology as almost any diagnosis that one could have, one could be status post. The same is true for "History of" and "Recurrent." Therefore, your terminology would start to grow rapidly, and a significant maintenance problem would occur whenever the definition of a diagnosis was altered. Expanding the reference terminology is more user-responsive if postcoordination is permitted. Frequently generated postcoordinations should be considered for inclusion in the terminology, as precoordinations, in subsequent versions of the terminology.

Although we may wish to say many things about CAD as a unit, there are still more granular ways to represent the same concepts. This similarity can be seen in many other constructions, for example, the combination of "large bowel" and "neoplasm, malignant" is equivalent to "colon cancer." This is particularly important for billing systems where the code for "colon cancer" might have a different ICD9-CM code than the two terms "large bowel" and "neoplasm, malignant."

One challenge in the development of a canonical vocabulary is to eliminate redundancy. Composition, while powerful, is also a source of considerable redundancy. The decision to allow postcoordination should be explicitly made by reference terminology designers, application designers, and implementers.

Methods for Creating Compositional Expressions

Previous Work by Other Groups

In 1999, Judith Wagner, Jeremy Rogers, Robert Baud, and Jean-Raoul Scherrer reported on the natural language generation of urologic procedures [7, 8]. Here they used a conceptual graph technique to apply translations for 172 rubrics from a common conceptual base between French, German, and English. They demonstrated that the GALEN model was capable of technically representing the concepts well; however, the language generation was often not presented in a form which native speakers of the target language would find natural. Trombert-Paviot et al. reported the results of the use of GALEN in mapping French procedures to an underlying concept representation [9]. Wroe et al. in 2001 reported the ability to integrate a separate ontology for drugs into the GALEN model [10]. Alan Rector in his expose "Clinical Terminology: Why is it so hard?" discusses the importance of, and ten most challenging impediments to, the development of compositional systems capable of representing the vast majority of clinical information in a comparable fashion [11]. In 2001, Professor Rector published one workable

method for integrating information models and terminology models [12].

Compositional Grammars

Compositional grammars are sets of rules/constraints which govern the creation of compositional expressions. Compositional grammars are an agreement between the authoring entity, the terminology development organization, and the receiving entity. That is to say, compositions above have a structural form that needs to be reliable. This forces certain constraints on the compositional system. First, compositions should be formed from semantics that are nonoverlapping (disjoint) and in the best case the description logic underpinning them should be completely descriptive so that the terminology can be normalized (see section "Normalization of both content and semantics"). Further, the grammar needs to specify a strict order of precedence for applying the semantics so, for example, the "has Finding Site" relation is applied before the "has Laterality" relation. The assignment of precedence to the semantics within a terminology will help to assure consistency in the compositional expressions.

Vocabulary-Based Strategies

Natural language processing is a complex computational task. Systems capable of understanding free speech are not presently available; however, many useful and reliable tools for the identification and manipulation of strings, lexical structures, and concepts have been developed [13–15]. Although the potential of NLP is not fully realized, we believe that harnessing all available information inherent in a free-text input string is a strategic goal for electronic medical records.

We advocate parsing free-text input strings into main concepts, qualifiers and modifiers and knowing the types of relationships that classes of Qualifiers can have with main concepts and each other. We can then provide better postcoordination of matched compound concepts (multiple Concept Unique Identifiers). Qualifiers are terms, which change the meaning of a term in a temporal or administrative sense, as opposed to a clinical sense (i.e., "History of," "Status/Post," "Recurrent," "Rule-Out," etc.) [16]. These compound concepts need to be linked/built-up in a meaningful and useful manner. Utilizing as much as possible the clues that we are given from the input string is an important mechanism for accomplishing this task. Generating the correct set of concepts for a given input phrase is the creation of Automated Compositional Expressions. Identifying the semantic dependency structure associated with these concepts is Automated Concept Dissection. An example of would be the input statement: "History of Benign Prostatic Hypertrophy BPH, status post transurethral resection of the prostate (TURP)."

Four unique concepts are represented in Fig. 5.1. We know that "History of" and "Status/Post" are both qualifiers and that BPH and TURP are both undifferentiated problems. The term "History of" can relate to just BPH or to both BPH and TURP. The term "Status/Post" always acts on the next concept or set of concepts and, therefore, must relate to TURP (S/P TURP). Hence, this expression could be interpreted as either (represented in ASN.1):

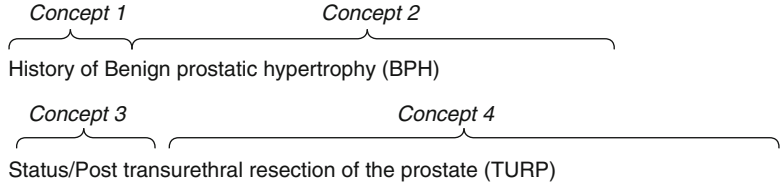


Fig. 5.1 Example of compositional segmentation of two clinical expressions

1. Concept {{Qualifier "Concept 1", Base-Concept {name "Concept 2"}}, {Qualifier "Concept 3", Base-Concept {name "Concept 4"}}} **or**
2. Concept {{Qualifier "Concept 1", Base-Concept {Concept {Base-Concept {name "Concept 2"}}, {qualifier "Concept 3", Base-Concept {name "Concept 4"}}}}}

In the first example, concept one qualifies just concept two, and in the second, it qualifies concepts two and four, whereas concept three always qualifies only concept four.

Compositional Expressions Using Multiple Terminologies

Many researchers in the associated fields of health data representation, knowledge representation, and terminologists believe that we will never have one reference terminology that will suit everyone's needs. It has been asserted that domain content experts who develop specialized reference terminologies take better care in the authoring of these reference terminologies than terminologists crafting a general health reference language. This is akin to the specialist model of health care which has permeated the practice of health care for over half a century. One such group of experts are the developers of the National Drug File – Reference Terminology (NDF-RT™) and its terminological cousin RxNorm (developed at the NLM and is included in the UMLS). The NDF-RT™ is a Veterans Administration (VA) led collaborative effort with the Food and Drug Administration (FDA), National Library of Medicine (NLM), the National Cancer Institute (NCI), and others. RxNorm is an NCVHS, and CHI endorsed standard representation of clinical drug and drug component names, and serves as an interlingua between drug systems' vendors proprietary code systems (e.g., FDB and Micromedex).

In some circumstances, it may be advantageous to combine two terminologies within one compositional expression. Perhaps, for example, we may want to use SNOMED CT for describing disorder-specific information, and we might wish to use the NDF-RT™ for medication-related content. In order to safely accomplish the construction of these expressions, we must understand the allowable semantics from each terminology and evaluate the suitability of the semantic relation to hold any concept from either source terminology in the operand or the specification position of the relationship triple (Operand–Relation–Specification) (see section "What is compositionality?"). An example of this would be linking indications for medication orders with the order itself.

Creating well-formed compositional expressions using concepts from different terminologies (e.g., NDF-RT™ and SNOMED CT) requires reconciliation of the overlap between the two terminologies in both content and semantics. The problem can be further broken down into a set of issues. (1) Overlapping content should not be used. (2) Overlapping semantics must either not be used or it must be formally defined where one uses each semantic within compositional expressions. (3) Rules must be developed regarding which linking semantics (e.g., description logic "roles") can be used to link concepts from one terminology to concepts from the other terminology.

The solution to problem (1) requires one to isolate the overlapping content between the two terminologies. One must identify hierarchies that contain common content between the two terminologies using expert review. To extend the NDF-RT™ and SNOMED CT example, expert review is required to identify hierarchies in SNOMED CT which contain NDF-RT™ or RxNorm concepts and vice versa. In cases where a substantial amount (≥80%) of the concepts in the hierarchy can be represented in the NDF-RT™ or RxNorm or their intent is identical (e.g., both terminologies have hierarchies of dose forms), one can try to eliminate the use of hierarchies which contain this overlapping information via the use of compositional grammars and compositional modeling "style guides." Compositional grammars are formal and a preferable method for expressing rules for the formation of compositional expressions. Style guides are less formal, and we believe should be directed at systems designers and implementers rather than end users, as the rules will likely be too complex for routine use by end users. Note that it is not acceptable to use the same concept coded in each

terminology even if it is identified as having originated from one of the terminologies as these concepts have different semantics and context within each of the component terminologies which in the end will lead to ambiguity. This ambiguity will lead to different retrieval sets depending on which terminologies explosions are used for the query. The table of excluded SNOMED CT concept IDs will enable you to filter out these concepts from the representational choices. The reviewers should perform the same process in the other direction (filtering out redundant material from the NDF-RT™) where those hierarchies are superior in SNOMED CT. Safe and effective content and semantics will be generated as the result of this process.

The solution to problem (2) requires one to first examine the allowable semantics from either terminology to identify any exact matches (both terminologies have "Isa" relationships). Next, you will need to identify overlapping semantics from linkage concepts within each terminology (e.g., "Has active ingredient" isa linkage concept in SNOMED CT where "Has-Active-Ingredient" is a semantic type in the NDF-RT™). Next, you need to identify partially overlapping semantics (e.g., "Has Component" in SNOMED CT and "Has Ingredient" in the NDF-RT™). Semantics that are partially overlapping can be used to create ambiguous compositional expressions and, therefore, need to be sanctioned as to their use in a regular and reliable manner. To accomplish this, you must specify in a compositional modeling style guide hierarchies and term sets which cannot serve as either the operand (the subject concept) or the specification (the target concept) for each particular partially overlapping semantic relation.

The solution to problem (3) is less straightforward than it might initially seem. Here we must understand the allowable semantics from each terminology and evaluate the suitability of the relation to hold any concept from either source terminology in the operand or the specification position of the relationship triple (Operand–Relation–Specification). This becomes more complicated, as in very complex expressions, the specification directly becomes the operand for the next relation in a potentially recursive fashion.

For example, if we take the expression (cellulitis of the left foot):

Cellulitis	Operand – Concept
Has_Finding_Site	Relation (or Role)
Left Foot	Specification – Concept
Left Foot	=> Becomes Operand for
Has_Laterality	Relation
Left	Specification – Concept

The decomposition of "cellulitis of the left foot" looks like this in SNOMED CT:

- Cellulitis (disorder) [128045006]
  - [has Finding Site]
    - Entire foot (body structure) [302545001]
      - [has Laterality]
        . Left (qualifier value) [7771000]

Using SNOMED CT with the NDF-RT™ to Create Composite Compositional Expressions

If we wanted to add a treatment for the condition "cellulitis of the left foot" using a medication specified in the NDF-RT™, we might need to form an expression such as:

- Cellulitis (disorder) [128045006]
  - [has Finding Site]
    - Entire foot (body structure) [302545001]
      - [has Laterality]
        . Left (qualifier value) [7771000]
  - [has Treatment]
    AMOXICILLIN TRIHYDRATE 500MG/CLAVULANATE K 125MG TAB [NDC: 00029608012] (drug) [C183848] [K]

Uses of composite compositional expressions using SNOMED CT and the NDF-RT™ include but are not limited to:

Steps in creating the environment that is capable of generating composite compositional expessions using these two terminologies include:

1. Eliminating overlapping content (see section "Errors associated with normalization of content")

   In our example above, Augmentin is represented in the NDF-RT™ as:

AMOXICILLIN TRIHYDRATE 500MG/CLAVULANATE K 125MG TAB [NDC: 00029608012] (drug) [C183848] [K]

While in SNOMED CT, it is represented as:

Co-amoxiclav 500 mg/125 mg tablet (product) [323539009] [K]

The compositional expression could be formed using either concept and be clinically correct. However, this would lead to semantic ambiguity, and this lack of specificity of meaning would lead to noncomparable data. Therefore, in this example, the SNOMED concept should be eliminated from the composite coding system in this example.

2. Eliminating overlapping semantics (see section "Errors associated with normalization of semantics")

   Using the same example, SNOMED CT has a semantic relation:

   Has active ingredient 127489000

   And the NDF-RT™ has a semantic relationship:

   Has_Ingredient

   Clearly, these two semantic relationships may have overlapping meaning. If we form a composition specifying a medication's ingredients, we want all instances of compositions that have the same meaning to have the same representation. For example:

   The Drug Augmentin has ingredients:

   Amoxicillin and Clavulanate both of which happen to be active ingredients and could clinically have been constructed/defined using either of these semantics. Therefore, a choice needs to me made to ensure interoperability.

3. Normalizing the content of the two terminologies (see section. "Normalization of both content and semantics")

   In SNOMED CT, we have the concept Cellulitis (disorder) [128045006]

   And in the NDF-RT™ we also have the concept Cellulitis (disease) [C1516] [K]

   But in SNOMED CT, cellulitis can also be defined as inflammation of the skin with etiology infectious disease. Since Inflammation (disease) [C3476] [K] and Infectious Diseases (disease) [C108] [K] are also concepts in the NDF-RT™, one might eliminate the directly duplicating concept Cellulitis (disease) [C1516] [K] and still be able to define Cellulitis using a compositional expression in NDF-RT™ concepts. Here you must find all uses of concepts that are being eliminated so that you do not break either source terminology. For example, in SNOMED CT, you have the concept Co-amoxiclav adverse reaction (disorder) [292985001] [K] which is defined using: Has Causative agent Amoxicillin with clavulanate potassium (product) [89519005] [K]. If you do not replace the NDF-RT™ concept for the SNOMED CT concept for cephalexin in the normative definition for Co-amoxiclav adverse reaction (disorder) [292985001] [K], then we break its description logic definition.

   This is essential because as you use the description logics from each terminology to normalize the content and semantics, the reasoner will not function properly if the definitional SNOMED CT code for Augmentin is not used in the compositional expression. For example, it would be possible to create a compositional expression using a combination of NDF-RT™ and SNOMED CT codes which would model:

   Adverse reactions (finding) [281647001] [K]
   HasEtiology
   AMOXICILLIN TRIHYDRATE 500MG/CLAVULANATE K 125MG TAB [NDC: 00029608012] (drug) [C183848] [K]

   This representation is incapable of being algorithmically normalized with the SNOMED CT concept Co-amoxiclav adverse reaction (disorder) [292985001] [K] given its current description logic definition.

Normalizing the Semantics of the Two Terminologies

This relates to defining which semantics can be used together within compositional expressions. For example, we might allow Has_Finding_Site and Has_Laterality from SNOMED CT to be combined with the Has_Treatment Relation bridging the NDF-RT™ concept with the

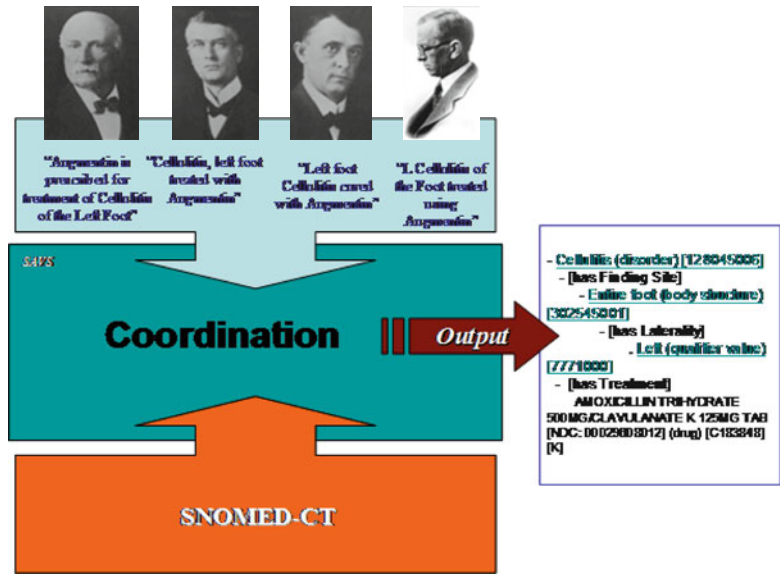


Fig. 5.2 A graphical depiction of the normalization of four disparate clinical expressions into a standard semantic representation

SNOMED CT expression (see Fig. 5.2). Here we define the semantic rules of interaction between the two terminologies.

4. Determining the rules for when to use each terminology (see section "Using SNOMED CT with the NDF-RT™ to create composite compositional expressions")

   These general rules will be employed to define which types of concepts will be represented by which terminology. This might be a rule-based expert system that knows that disease descriptions or findings might come from SNOMED CT where drug names may come from the NDF-RT™.

5. Determining the syntax and semantics for representing the bridge between the two terminologies (see section "Using SNOMED CT with the NDF-RT™ to create composite compositional expressions")

   This is where we define the semantic and syntactic structures needed for interoperability. The example below is one suggested semantic description. HL7 Templates might be a suggested method for combining a standard syntax with this semantic representation:

   - Cellulitis (disorder) [128045006]
     - [has Finding Site]
       - Entire foot (body structure) [302545001]
         - [has Laterality]
           . Left (qualifier value) [7771000]
     - [has Treatment]
       AMOXICILLIN TRIHYDRATE 500MG/CLAVULANATE K 125MG TAB [NDC: 00029608012] (drug) [C183848] [K]

6. Applying the rules generated in point six consistently for both the storage and retrieval of information (see Sect. 10.2.1.2)

   This makes the point that the compositional rules combined with the base terminological representation are needed for true data interoperability. As the rules imbue the representation with stronger semantic interoperability, they are just as important for data retrieval as they are for data storage.

Example of Storage and Retrieval of Codified Data

The process of entry and retrieval of codified data should not require the user to understand the codes or structure of the terminology that encodes their data. The workflow (see Fig. 5.3) should provide a sensitive and specific retrieval set

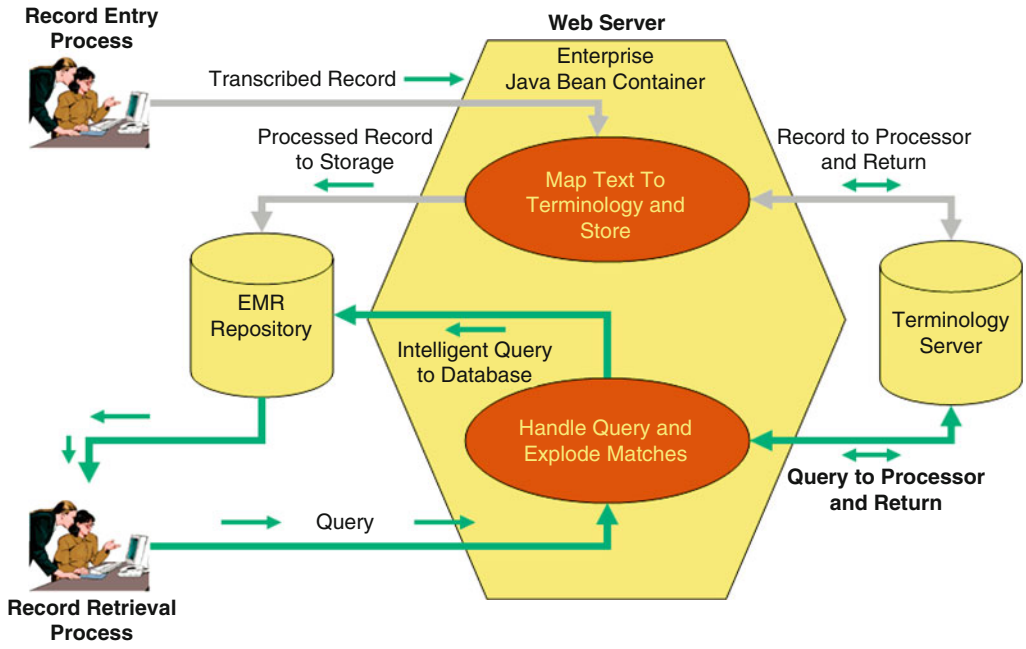


Fig. 5.3 Workflow for data entry and retrieval of information where the user dictates in their native language and the system uses a terminology server to code the data. When they query this same dataset, the information is returned by parsing the user's critical question which the system codifies and matches using the codes against the stored data

without the user having to be aware of how that is made to happen. The valid relationships available from within SNOMED CT from its release one are shown in Table 5.2.

Formal Knowledge Representation

All of these functions require a method for formal knowledge representation. Description logic is a subset of first-order logic used to rigorously define concepts within a terminology. Defining concepts "formally" and explicitly, i.e., by specifying their interrelationships, offers many advantages to terminology construction, maintenance, and quality assurance. Most importantly, description logics facilitate terminology life cycle automation that improves quality and reduces cost.

Using OWL as a Formal Language to Represent Knowledge Unambiguously

The Web Ontology Language (OWL) description logic has constructs to support composition as described above. OWL is the official language of the semantic web and as such will be the de facto method of communicating semantics over the World Wide Web.

OWL is a proper constraint language which handles decidably the core of first-order predicate logic [17]. OWL is an outgrowth of the DAML+OIL project with the following exceptions:

- The removal of qualified number restrictions, per a decision of WebOnt
- The ability to directly state that properties can be symmetric, per a decision of WebOnt
- The absence in abstract syntax of some abnormal DAML+OIL constructs. Particularly

Table 5.2 Valid relationship type concepts in second release of SNOMED CT

Name and Concept Id
Access 260507000
Access instrument 370127007
Approach 260669005
Associated etiologic finding 363715002
Associated finding 246090004
Associated function 116683001
Associated morphology 116676008
Causative agent 246075003
Communication with wound 263535000
Component 246093002
Course 260908002
Direct device 363699004
Direct morphology 363700003
Direct substance 363701004
Episodicity 246456000
Finding site 363698007
Has active ingredient 127489000
Has definitional manifestation 363705008
Has focus 363702006
Has intent 363703001
Has interpretation 363713009
Has specimen 116686009
Indirect device 363710007
Indirect morphology 363709002
Interprets 363714003
Laterality 272741003
Location 246267002
Measures 367346004
Method 260686004
Occurrence 246454002
Onset 246100006
Part of 123005000
Pathological process 370135005
Priority 260870009
Procedure site 363704007
Recipient category 370131001
Revision status 246513007
Severity 246112005
Stage 258214002
Subject of information 131195008
Temporally follows 363708005
Using 261583007

restrictions with extra components, and a difference in their meaning

The only difference between full OWL and OWL DL is that names for classes, individuals, properties, and datatypes must be disjoint (i.e., no name should occur in more than one of these categories). This restriction assures that only first-order logic is represented and makes reasoning over these ontological constructions less complex (they practically will finish their queries in less than polynomial time). OWL Lite was not chosen, as it cannot support cardinalities of greater than one. This restriction is not applicable to OWL DL or OWL Full. See Appendix for further details of OWL.

Safe Compositional Expressions

Definition

A safe compositional expression is one that is logically well formed and is capable of being normalized.

To be logically well formed, an expression must have the syntax and semantics consistent with first-order logic and must also use a set of legal semantics that have a defined set of Operands and Specifications (i.e., Operand–Relation–Specification; see section "What is compositionality?" for an example).

Normalization of Both Content and Semantics

Creating safe compositional expressions requires the normalization of both content and semantics. Normalization is a separate but important terminological issue which affects grammars. Normalization is defined as the ability to identify every representational format that confers the same meaning as being equivalent (i.e., unambiguous representation). For example, in SNOMED CT v1.0, cellulitis is defined as has_Finding_Site Skin and has_Morphology Cellulitis. Morphology cellulitis is defined as Isa Inflammation. Dermatitis is defined as has_Finding_Site Skin and has_Morphology Inflammation. Algorithmically, from a classification mechanism using a description logic classifier, these two entities are defined identically and therefore would not be able to be easily

separated by the terminologies classifier. This means that an attempt to normalize the terminology would suggest that cellulitis and dermatitis are in fact the same disorder. In truth, we know that this is not the case. So where did SNOMED go wrong. At a minimum, they left out the has_Etiology Infectious Disease from the description logic definition of cellulitis.

Examples of applicable semantics are Is_a relations and Part_of relations (per Alan Rector there are seven types of partonomy). Other clinically relevant semantics can be defined as needed by the terminological application. This is valid as long as one remembers the original assumptions of the underlying logic which are (1) that the relations are independent from one another (disjoint) and (2) that for any domain the relations are sufficient to provide distinct definitions for every concept within the domain.

In a previous study, we delineated the combined semantics from ICD9-CM and SNOMED RT as a component of the necessary work to make these two terminologies interoperable [18]. The overall method involved four steps. First, we used SuperTagger, a natural language processing (NLP) parser, to identify the verbs and verb phrases in ICD9-CM textual descriptions. Second, we manually reviewed the identified phrases as candidate semantic relations. Third, verb phrases determined to be valid semantic relations were manually ordered into an ontology and were matched to the SNOMED RT version 1.0 modifier/linkage hierarchy. Those without an exact match in the SNOMED RT modifier/linkage hierarchy were added. The result of the first three steps is a merged ontology of SNOMED RT description logic relations and newly discovered semantic relations that had been implicit in ICD9-CM.[1] Finally, we used existing tools to identify implicit occurrences of merged set of semantics within SNOMED RT and to create alternative representations with explicit semantics. Because the methodology of this previous study is directly relevant to the topic of this white paper, further details of the methodology are given below.

Example of Normalizing Semantics

Identifying the Implied Semantics from ICD9-CM

In the article noted above, a full text version of the ICD9-CM codes and their textual descriptions was obtained from HSS, Inc {Hamden, Connecticut}. ICD9-CM textual descriptions were processed using SuperTagger, a freely downloadable natural language parser available from the University of Pennsylvania [19]. SuperTagger was used to identify the parts of speech of all terms in ICD9-CM textual descriptions (e.g., nouns, verbs, noun phrases, verb phrases). Manual review was used to eliminate mappings without a verb form assigned.

All of the verbs or verb phrases were reviewed as candidate semantic relations by at least one author. A second author was engaged to discuss the applicability of a verb phrase as a candidate semantic relation in cases where questions arose. Verb phrases were manually reviewed and organized using their context within the ICD9 terms as clues as to their usage. For example, we determined that the phrase "preexisting condition" was used in a similar manner to "has_History_of" in terms such as "gastric perforation with preexisting peptic ulcer disease." The reproducibility of the manual review process was not evaluated.

We added each of the newly identified semantics to the modifier/linkage hierarchy of SNOMED RT if it was not already present. We manually assigned synonymy to terms that appeared to be used in the same fashion within ICD9-CM and the SNOMED RT linkage hierarchy. Finally, we extended the ICD9 ontology of semantics described above to cover the merged set of semantics and associated synonyms (see Table 5.3).

Errors Associated with the Use of Compositionality

Errors of composition fall into three basic categories. The first is error associated with normalization of the content of the terminology.

[1]By implicit we mean relations that were present in the ICD9 textual descriptions

Table 5.3 Combined ontology of semantics, from SNOMED RT and ICD9-CM21

Merged ontology	Found in	Synonyms
1. Has-Etiology	Both	Secondary to, caused by, due to, arising from, resulting from, from other, referable to, of underlying, cause, In, arising in
1.1 Caused by other than	ICD	
1.2 Transmitted by	RT	
2. Without	Both	Specified as excluding, without mention of, lack of, free of
3. Has-Complication	Both	Complicated by, complications of, secondary, effect of, induced, effecting, resulting in, interfering with, causing, has development
3.1 Late-Complication	ICD	Late effect of, residual effect
3.2 Early-Complication	ICD	Early effect of, initial effect
3.3 Drug induced	ICD	Drug effect
3.3.1 Steroid induced	ICD	
3.4 Metastatic to	RT	
4 Has-Involvement	Both	Involving, including, not free of
4.1 Extending	ICD/RT	Extending into, extension of
5. Has-Association	Both	Associated with, with, states association with, mention of, with mention, occurring in
5.1 With type	RT	
5.2 With shape	RT	
5.3 With staging	RT	
5.4 With pattern	RT	
5.5 With color	RT	
5.6 Bounded by	RT	
5.7 With frequency	RT	
5.8 With size	RT	
5.9 With distance	RT	
5.10 With laterality	RT	
5.11 With odor	RT	
6. Has-History of	ICD	Preexisting
7. Or	ICD	Or other
8. Has-Confirmation	ICD	Confirmed by, found by, found
9. Not	ICD	Not found
9.1 Except for	RT	
10. Has-Specification	Both	Specified form, with other specified, including, other specified, specified as
11 Has-Action	ICD	Similarly acting, related acting
12 Has-Temporal Relationship	ICD	
12.1 Before	ICD	Before onset of
12.2 During	ICD	
12.2.1 Onset of	ICD	Initiating, appearance of, onset
12.2.1.1 Rapid onset of	ICD	Rapid onset
12.2.1.2 Insidious onset of	ICD	Insidious onset
12.2.2 Maintenance of	ICD	Maintaining, maintenance
12.2.3 Late	ICD	
12.3 Status Post	Both	s/p, Post-, after, following, After, followed by
12.3.1 Resolved	ICD	Resolution, state of resolution
13 Has-Resistance	ICD	Resistant to
13.1 Drug resistant	ICD	
14 Treated with	RT	With treatment, with therapy, treated by
14.1 Controlled by	RT	Controlled with
14.1.1 Well controlled by	RT	well controlled with, well controlled by

Table 5.2 combines the semantics of SNOMED RT with the semantics of ICD9-CM. This table represents the minimal set of semantics that would need to be modeled in SNOMED to map to ICD9 algorithmically

The second is error associated with normalization of the semantics of the terminology, and the third is errors that result from ambiguous or misleading interaction between the presentation of the compositional terminology and clinicians.

In order to understand errors of composition, one must first understand what a compositional system is and what it is not. A compositional system should allow users to postcoordinate concepts in order to represent a more specific notion. Postcoordination implies the ability of a user or system to put combinations of concepts together to form a compositional expression, which did not exist within the terminology to describe this more specific notion. For example, a clinician might describe a new onset cardiac event as:

"Myocardial Infarction"
  Has-Acuity "Acute"
  Has-Location "Anterolateral Wall"

This composition must be constructed using rules for composition that protect against unrecognized ambiguity. Rules can be created and documented via compositional grammars (section "Compositional grammars") and by compositional modeling "style guides." Normalization implies the ability to recognize all representations that express the same meaning as being algorithmically equivalent.

Errors Associated with Normalization of Content

Normalization of content is accomplished when all possible representations of the same concept using the same semantics are identifiable (algorithmically) as being equivalent. Taking the example above, one would need to be able to recognize as equivalent the following additional forms:

"Acute Myocardial Infarction"
  Has-Location "Anterolateral Wall"

"Myocardial Infarction, Anterolateral Wall"
  Has-Acuity "Acute"

"Acute Myocardial Infarction, Anterolateral Wall"

"Myocardial Infarction, Anterolateral Wall, Acute"

"Myocardial Infarction"
  Has-Location "Anterolateral Wall"
  Has-Acuity "Acute"

"Infarction"
  Has-Location "Anterolateral Wall of Myocardium"
  Has-Acuity "Acute"

In order to accomplish this goal, one requires a compositional terminological system which is (a) description logic based, (b) where all atomic and precoordinated concepts are fully defined in the terminological system, (c) where the rules of composition handle all types of composite variations that are allowable (can be created using the system). Please note that fully defined concepts are autoclassifiable by the description logic classifier.

Errors Associated with Normalization of Semantics

Errors associated with the normalization of semantics are harder to recognize than normalization of content during design of a terminological system. Nonnormalized semantics have overlapping meaning, which leads to unrecognized ambiguity. Unrecognized ambiguity results when you can create two or more compositional expressions that have the same clinical meaning but have different representations that cannot be determined to be equivalent algorithmically. For example, one could represent a "Laproscopic Cholecystectomy" either as:

"Surgical Procedure: Excision"{Has Site Gallbladder}, {Has Method Endoscopic}
or
"Surgical Procedure: Excision" {Has Site Gallbladder}, {Using Device Endoscope}.

This ambiguous representation is possible as the semantics "Has-Method" and "Using-Device" are partially overlapping because performing a procedure using a device implies one or more methods (e.g., the "endoscopic" method). In this example, ambiguous representations can be avoided by using rules of composition documented in a compositional modeling "style

guide," or by careful construction of partially overlapping semantics. In this circumstance, the simplest solution is to create a "style guide" rule that stipulates when each semantic ("has_method" and "using_device") is to be used. A more complicated solution is to use formal role subsumption to define the relationship between the overlapping semantics, i.e., that using a device to perform a procedure "is-a" type of method.

Errors in Interfacing with Clinicians

Even after we go to all the trouble of normalizing the content and semantics of the terminology, we still may create ambiguous data. Clinicians come to a compositional system with a long history of training and have specific experiences and knowledge that colors the meanings that they associate with a particular representation. One might ask "Is the cause already lost?" I would answer "No" as clinicians have a remarkable adaptive ability which will allow them to understand how any consistently presented system works.

However, it is incumbent upon systems developers to follow a few important rules. One, developers must try as much as possible to present similar information to users in a consistent fashion. Users will get used to a consistent style and will allow them to more frequently recognize complex compositional constructs completely. Two, developers must understand their users' environment so that the form of composition meets the needs of the user. Unneeded complexity can turn off the casual user of such a system. Three, follow user-centered design principles in constructing your user interfaces. When possible, employ formal usability testing to ensure that your application is understandable and usable.

Architectures to Implement Compositionality

Distributed computing theory recognizes that servers are efficient at handling large amounts of information. They, however, are not good at handling process intensive tasks, by virtue of the fact that many users will in all likelihood be using the server simultaneously. Therefore, we recommend pushing process intense tasks to the client when feasible, given the availability of relatively cheap cycle time.

Server Based Strategies

Vocabulary Storage and Retrieval

The capability of massive data storage and retrieval with buffering of indices, which can be accessed by multiple simultaneous processes, makes server side retrievals fast and efficient. Maintenance and updating of the vocabulary need be done in only one place for all users to benefit. Better version consistency can be maintained.

Universal Unique Identifiers (UUID) for Compound Concept Unique Identifiers

We will never want to maintain a concept in the base vocabulary for every compound concept that a user may want to express. For example, "History of BPH s/p TURP" does not make most workers' lists of atomic concepts. On the other hand, a clinician may very well wish to make this statement regarding one of their patients. Each and every time such a reference is used, we would want to capture its meaning, and if another clinician wrote "Hx of BPH two years after a TURP," it would be nice if a system could recognize these as being related to the same set of concepts. This requires that the server serve up the same identifier not only for unique concepts but also for unique compound concepts. We have used these in several different ways. First, by presenting them to users from a search, we encourage the identical construction of a compositional expression for a given term rather than a slight variation that may be clinically reasonable but not identical for retrieval purposes. Secondly, we use them to avoid adding new concepts to the terminology. If a new idea crops up in medicine, we create a composition to describe it and use the native coding system to represent the concept. If and when a code becomes available for this, we can mark the compositional expression obsolete as of a certain date and give a pointer to the new concept.

Upon retrieval, we will know that before a certain date, this notion can be retrieved using the compositional expression, and after that date, it will be retrievable using the new concept id.

Making the Most out of Your Retrieval List

We make use of simple rules of composition, which uses ontology of qualifiers that can be combined in selected ways with other concepts selected by UMLS semantic type (i.e., problem, disorder, etc.). Multiple qualifiers can be combined with multiple other concepts to provide a short list of retrievals, which a clinician might choose as their problem list entry. These postcoordinated terms are presented at the top of the retrieval list, but with no indication that they differ from any other term presented on the retrieval list (e.g., unique atomic concepts, unique precoordinated concepts).

Conclusions

Compositionality is the ability of a vocabulary system to represent nonatomic expressions in an unambiguous way. In this chapter, we defined the types of composition, which can occur and differentiated the terms precoordination, postcoordination, and user-directed coordination. We have discussed the risks and benefits of using a compositional system, and we have suggested methods that minimize the risk of creating ambiguous representations using compositional expressions. We presented a tractable method for using more than one terminology in creating compositional expressions. We have discussed normalization of both content and semantics and have shown how this process leads to comparable data. We have discussed the syntax, semantics, and content of compositional expressions and provided examples. We have provided a stepwise protocol for the analysis of two terminologies to determine their degree of compatibility. Then we have suggested a method for determining the content and semantics from each terminology that can be safely used in creating composite compositional expressions. In specific, we looked at how three terminologies SNOMED CT, ICD9-CM, and the NDF-RT™ can be used together. This included a method to identify overlapping content and semantics. This method is also extensible to the LOINC terminology. Further, we have suggested additional semantics, which if added to these reference terminologies would lead to more accurate data representation and therefore an improved level of interoperability. Lastly, we defined the logic for support of formal web-based representation of interoperable data using RDFS XML in OWL, which is the description logic language of the semantic web.

Compositionality is important. The expressivity associated with a compositional system is significantly greater than using the same system without composition (51% vs. 92.3%; $p < 0.001$ when considering SNOMED CT for clinical problem statements). The use of composition comes with the risk of creating unwanted ambiguous expressions. The path to mitigating this risk is to (a) only use semantics that are independent (nonoverlapping in meaning or formally defined with subsumption relationships) (b) make sure that within the terminology the semantics are instantiated everywhere that they apply (c) that the formalism for creating compositional expressions is logically rigorous and is applied in a consistent way based on a formal set of rules or on a compositional modeling "style guide" (d) that there is a process for normalizing the terminology (identifying all of the different compositions that would have the same meaning) (e) that there is a classifier associated with the formal terminology so that new terms can find their appropriate place in the ontology. Associated with these basic principles of composition is the need for a terminology server that is capable of automatic generation of compositional expressions when provided with a textual input, for both information capture and retrieval.

Using two or more terminologies together in the same compositional expression requires that all of the overlapping content and semantics be identified and that rules be created to define when to sanction the use of

each terminologies content and semantics in the construction of compositional expressions using the compositional system.

The use of a properly formed compositional system will allow you to turn your text into data that data into information, the information can be aggregated into knowledge and agents can use that knowledge as intelligence. Although some significant challenges remain, we as healthcare informaticians are now poised to represent a large and important set of health data specifically and exactly by using a compositional system of content and semantics.

Questions

1. What is compositionality?
   (a) The ability to compose single concepts within a terminology
   (b) Compositions are expressions made up of more than one concept where all pairs are joined by relationships
   (c) Are expressions not found in the original terminology
   (d) Are precoordinated concepts
2. Acceptable formalisms for defining compositional expressions are?
   (a) OWL
   (b) Conceptual Graphs
   (c) Description logics
   (d) DAGs
   (e) All of the Above
3. How many concepts does it take to represent billions of clinical utterances using a compositional terminology?
   (a) <106
   (b) <1010
   (c) <109
   (d) <108
   (e) <107
4. How many concepts does it take to represent billions of clinical utterances using a noncompositional terminology?
   (a) <106
   (b) <1010
   (c) <109
   (d) <108
   (e) <107
5. The Operand in a compositional expression is?
   (a) The operator
   (b) The relationship
   (c) The main concept of the expression
   (d) The refining characteristic
6. The Specification of the compositional expression is?
   (a) The operator
   (b) The relationship
   (c) The main concept of the expression
   (d) The refining characteristic
7. Having independent semantics means?
   (a) That only one semantic can be used in a compositional expression
   (b) That the semantics used are nonoverlapping in meaning
   (c) That the concepts are nonoverlapping in meaning with the semantics
   (d) That the semantics are defined by terms in the terminology
8. Uniformity of relations means?
   (a) That the representational form of the relations is consistent
   (b) That uniformity is used in assigning names to relations
   (c) That relations are linked to concepts
   (d) That relations are applied everywhere they are applicable within the terminology
9. Logical consistency means?
   (a) That you use the same logic throughout the development of the terminology
   (b) That you use the same logic for the use of postcoordinated compositional expressions
   (c) That you use the same logic for the development and use of the terminology
   (d) That you use the same logic as a DL classifier and a Conceptual Graph Classifier
10. Semantic normalization is present when you are able to?
    (a) Identify all of the precoordinated and postcoordinated concepts with the same meaning
    (b) Identify all the precoordinated concepts with the same meaning as an atomic concept
    (c) Identify all the atomic concepts with the same meaning as a precoordinated concept
    (d) Identify all the atomic concepts with the same meaning as a postcoordinated concept

11. Computational normalization is when the classifier used to build the terminology is capable of auditing the terminology and its postcoordinated expressions?
    (a) True
    (b) False
    (c) Unknown
    (d) Unknowable
12. Colloquial normalization is when?
    (a) Things that are commonly said are normalized
    (b) Normalization is used when it is practical to do so
    (c) Allowable grammars are limited to what can be normalized by the classifier
    (d) Normalization is limited to semantics that have worked in the past
13. It is worthwhile creating a precoordinated concept in a terminology when?
    (a) Experts think you should
    (b) The terminology looks more clinical with the concept
    (c) The terminology looks incomplete without the concept
    (d) You need to associate the concepts with other information in the terminology
14. Creating a precoordinated compositional expression?
    (a) Adds new knowledge to the terminology
    (b) Adds more information to the terminology
    (c) Does not add new knowledge to the terminology
    (d) Adds semantics to the terminology
15. When do you want to postcoordinate concepts?
    (a) When the specification would apply to a large number of concepts
    (b) When the expression uses semantics which have high reliability of assignment
    (c) When forming an expression for "History of CABG"
    (d) All of the above
16. When using more than one terminology to form compositional expressions?
    (a) It is important to delete all overlapping contents from the larger terminology
    (b) It is important to normalize the content and semantics of each terminology
    (c) It is important to delete all the overlapping semantics from the larger terminology
    (d) It is important to delete the overlapping content from the smaller terminology
17. When using multiple terminologies in compositional expressions?
    (a) It is important to examine the concepts from both terminologies
    (b) It is important to map together the concepts from each terminology together
    (c) It is important to map together the concepts and their description logic references
    (d) It is important not to be fooled by the name description logic
18. Composition improves SNOMED CT's coverage of clinical problems by?
    (a) 21%
    (b) 31%
    (c) 41%
    (d) 51%
19. Compositionality increases terminological expressivity by?
    (a) 10^5
    (b) 10^4
    (c) 10^3
    (d) 10^2
20. A safe compositional expression is one that?
    (a) Is logically well formed and is capable of being normalized
    (b) Has already been normalized with the terminology
    (c) Has been written using a description logic representation
    (d) Has no collisions with precoordinated concepts from the terminology

Appendix

Addendum Details of OWL

OWL Class Axioms

The full abstract syntax has more-general versions of the OWL Lite class axioms where superclasses, more-general restrictions, and Boolean combinations of these are allowed. Together, these constructs are called descriptions.

<axiom> ::= Class(<classID><modality>{<annotation>} {<description>})
<modality> ::= complete | partial

In the full abstract syntax, it is also possible to make a class exactly consist of a certain set of individuals, as follows.

<axiom> ::= EnumeratedClass
(<classID>{<annotation>} {<individualID>})

Finally, in the full abstract syntax, it is possible to require that a collection of descriptions be pairwise disjoint, or have the same members, or that one description is a subclass of another. Note that the last two of these axioms generalize the first kind of class axiom just above.

<axiom> ::= DisjointClasses
(<description>{<description>})
<axiom> ::= EquivalentClasses
(<description>{<description>})
<axiom> ::= SubClassOf
(<description><description>)

OWL Descriptions

<axiom> ::= Class(<classID><modality> {<annotation>} {<super>})
<modality> ::= complete | partial
<super> ::=<classID> | <restriction>

In OWL Lite, it is possible to state that two classes are the same.

<axiom> ::= EquivalentClasses
(<classID>{<classID>})

Descriptions in the full abstract syntax include class IDs and the restriction constructor. Descriptions can also be Boolean combinations of other descriptions, and sets of individuals.

<description> ::= <classID>
| <restriction>
| unionOf({<description>})
| intersectionOf({<description>})
| complementOf(<description>)
| oneOf({<individualID>})

OWL Restrictions

<restriction> ::= restriction(<datavaluedPropertyID> {allValuesFrom(<datatypeID>)}
{someValuesFrom(<datatypeID>)}
[<cardinality>])
<restriction> ::= restriction(<individualvaluedPropertyID> {allValuesFrom(<classID>)}
{someValuesFrom(<classID>)}
[<cardinality>])
<cardinality> ::= minCardinality(0) | minCardinality(1) |
| maxCardinality(0) |
maxCardinality(1) |
| cardinality(0) | cardinality(1)

Restrictions in the full abstract syntax generalize OWL Lite restrictions by allowing descriptions where classes are allowed in OWL Lite and allowing sets of data values as well as datatypes. The combination of datatypes and sets of data values is called a data range. In the full abstract syntax, values can also be given for properties in classes. As well, cardinalities are not restricted to only 0 and 1.

<restriction> ::= restriction(<datavaluedPropertyID>{allValuesFrom(<dataRange>)}
{someValuesFrom(<dataRange>)}
{value(<dataLiteral>)}
{<cardinality>})
<restriction> ::= restriction(<individualvaluedPropertyID>{allValuesFrom(<description>)}
{someValuesFrom(<description>)}
{value(<individualID>)}
{<cardinality>})
<cardinality> ::= minCardinality(<non-negative-integer>)
| maxCardinality(<non-negative-integer>)
| cardinality(<non-negative-integer>)

A dataRange, used as the range of a datavalued property and in other places in the full abstract syntax, is either a datatype or a set of data values.

<dataRange> ::= <datatypeID>
<dataRange> ::= oneOf({<typedDataLiteral>})

As in OWL Lite, there is a side condition that properties that are transitive, or that have transitive subproperties, may not have cardinality conditions expressed on them in restrictions.

OWL Property Axioms

<axiom> ::= DatatypeProperty (<datavaluedPropertyID>{<annotation>}
{super(<datavaluedPropertyID>)}

{domain(<classID>)}
{range(<datatypeID>)}
[Functional])

<axiom> ::= ObjectProperty (<individualvaluedPropertyID>{<annotation>} {super(<individualvaluedPropertyID>)}
{domain(<classID>)}
{range(<classID>)}
[inverseOf(<individualvaluedPropertyID>)] [Symmetric]
[Functional | InverseFunctional | Functional InverseFunctional | Transitive])

The following axioms make several properties be the same, or make one property be a subproperty of another.

<axiom> ::= EquivalentProperties(<datavaluedPropertyID>{<datavaluedPropertyID>})

<axiom> ::= SubPropertyOf(<datavaluedPropertyID><datavaluedPropertyID>)

<axiom> ::= EquivalentProperties(<individualvaluedPropertyID>{<individualvaluedPropertyID>})

<axiom> ::= SubPropertyOf(<individualvaluedPropertyID><individualvaluedPropertyID>)

Property axioms in the full abstract syntax generalize OWL Lite property axioms by allowing descriptions in place of classes and data ranges in place of datatypes in domains and ranges.

<axiom> ::= DatatypeProperty (<datavaluedPropertyID>{<annotation>}
{super(<datavaluedPropertyID>)}
{domain(<description>)}
{range(<dataRange>)}
[Functional])

<axiom> ::= ObjectProperty
(<individualvaluedPropertyID>
{<annotation>} {super(<individualvaluedPropertyID>)}
{super(<individualvaluedPropertyID>)}
{domain(<description>)}
{range(<description>)}
[inverseOf(<individualvaluedPropertyID>)] [Symmetric]
[Functional | InverseFunctional | Functional InverseFunctional | Transitive])

OWL Language and Its Formal Logic

The semantics here start with the notion of a vocabulary, which can be thought of as the URI references that are of interest in a knowledge base. It is, however, not necessary that a vocabulary consist only of the URI references in a knowledge base.

An OWL vocabulary V is a set of URI references, including owl:Thing, owl:Nothing, and rdfs:Literal. Each OWL vocabulary also includes URI references for each of the XML schema nonlist built-in simple datatypes. In the semantics, LV is the (nondisjoint) union of the value spaces of these datatypes.

An Abstract OWL interpretation with vocabulary V is a four-tuple of the form: I=<R, S, EC, ER>

where

- R is a nonempty set of resources, disjoint from LV
- S : V → R
- EC : V → 2^R ∪ 2^LV
- ER : V → 2^(R × R) ∪ 2^(R × LV)

S provides meaning for URI references that are used to denote OWL individuals, while EC and ER provide meaning for URI references that are used as OWL classes and OWL properties, respectively.

Abstract OWL interpretations have the following conditions having to do with datatypes:

1. If d is the URI reference for an XML schema non-list built-in simple datatype, then EC(d) is the value space of this datatype.
2. If c is not the URI reference for any XML schema non-list built-in simple datatype, then EC(c) is a subset of R.
3. If d,l is a datatype,literal pair, then D(d,l) is the data value for l in XML schema datatype d.

EC is extended to the syntactic constructs of <description>s, <dataRange>s, <individual>s, and <propertyValue>s as follows:

Syntax - S	**EC(S)**
owl:Thing	R
owl:Nothing	{ }
rdfs:Literal	LV
complementOf(c)	R - EC(c)
unionOf(c_1 … c_n)	$EC(c_1) \cup \ldots \cup EC(c_n)$
intersectionOf(c_1 … c_n)	$EC(c_1) \cap \ldots \cap EC(c_n)$
oneOf(i_1 … i_n)	$\{S(i_1), \ldots, S(i_n)\}$
oneOf(d_1,l_1 … d_n,l_n)	$\{D(d_1,l_1), \ldots, D(d_n,l_n)\}$
restriction(p x_1 … x_n)	EC(restriction(p x_1))∩… ∩EC(restriction(p x_n))
restriction(p allValuesFrom(r))	$\{x \in R \mid <x,y> \in ER(p) \rightarrow y$ EC(r)}
restriction(p someValuesFrom(e))	$\{x \in R \mid \in <x,y>$ $\in ER(p) \rightarrow y \in EC(e)\}$
restriction(p value(i))	$\{x \in R \mid <x,S(i)> \in ER(p)\}$
restriction(p value(d,l))	$\{x \in R \mid <x,D(d,l)>$ $\in ER(p)\}$
restriction(p minCardinality(n))	$\{x \in R \mid card(\{y : <x,y>$ ER(p)$\}) \leq n\}$
restriction(p maxCardinality(n))	$\{x \in R \mid card(\{y : <x,y>$ ER(p)$\}) \geq n\}$
restriction(p cardinality(n))	$\{x \in R \mid card(\{y : <x,y>$ ER(p)$\}) = n\}$
Individual(annotation(…) … annotation(…) type(c_1) … type(c_m) pv_1 … pv_n)	$EC(c_1) \cap \ldots \cap EC(c_m) \cap$ $EC(pv(pv_1)) \cap \ldots \cap$ $EC(pv(pv_n))$
Individual(i annotation(…) … annotation(…) type(c_1) … type(c_m) pv_1 … pv_n)	$\{S(i)\} \cap EC(c) \cap \ldots \cap$ $EC(c_m) \cap EC(pv(pv_1))$ $\cap \ldots \cap EC(pv(pv_n))$
pv(p Individual(…))	$\{x \in R \mid$ $\exists y \in EC(Individual(\ldots)) :$ $<x,y> \in ER(p)\}$
pv(p id), for id an individualID	$\{x \in R \mid <x,S(id)> \in ER(p)$ }
pv(p d,l)	$\{x \in R \mid <x,D(d,l)> \in ER(p)$ }

An Abstract OWL interpretation, I, is an interpretation of OWL axioms and facts as given in the table below. In the table, optional parts of axioms and facts are given in square brackets ([…]) and have corresponding optional conditions, also given in square brackets.

Directive	**Conditions on interpretations**
Class(c complete annotation(…) … annotation(…) $descr_1$ … $descr_n$)	$EC(c) = EC(descr_1) \cap \ldots \cap$ $EC(descr_n)$
Class(c partial annotation(…) … annotation(…) $descr_1$ … $descr_n$)	$EC(c) \subseteq EC(descr_1) \cap \ldots \cap$ $EC(descr_n)$
EnumeratedClass(c annotation(…) … annotation(…) i_1 … i_n)	$EC(c) = \{ S(i_1), \ldots, S(i_n) \}$
DisjointClasses(d_1 … d_n)	$EC(d_i) \cap EC(d_j) = \{ \}$ for $1 \leq i < j \leq n$
EquivalentClasses(d_1 … d_n)	$EC(d_i) = EC(d_j)$ for $1 \leq i < j \leq n$
SubClassOf(d_1 d_2)	$EC(d_1) \subseteq EC(d_2)$
DataProperty(p annotation(…) … annotation(…) super(s_1) … super(s_n) domain(d_1) … domain(d_n) range(r_1) … range(r_n) [Functional])	$ER(p) \subseteq ER(s_1) \cap \ldots \cap$ $ER(s_n) \cap EC(d_1) \times LV$ $\cap \ldots \cap EC(d_n) \times LV \cap$ $R \times EC(r_1) \cap \ldots \cap$ $R \times EC(r_n)$ [ER(p) is functional]
IndividualProperty(p annotation(…) … annotation(…) super(s_1) … super(s_n) domain(d_1) … domain(d_n) range(r_1) … range(r_n) [inverse(i)] [Symmetric] [Functional] [InverseFunctional] [Transitive])	$ER(p) \subseteq ER(s_1) \cap \ldots \cap$ $ER(s_n) \cap EC(d_1) \times R$ $\cap \ldots \cap EC(d_n) \times R \cap$ $R \times EC(r_1) \cap \ldots \cap$ $R \times EC(r_n)$ [ER(p) is the inverse of ER(i)] [ER(p) is symmetric] [ER(p) is functional] [ER(p) is inverse functional] [ER(p) is transitive]
EquivalentProperties(p_1 … p_n)	$ER(p_i) = ER(p_j)$ for $1 \leq i < j \leq n$
SubPropertyOf(p_1 p_2)	$ER(p_1) \subseteq ER(p_2)$
SameIndividual(i_1 … i_n)	$S(i_j) = S(i_k)$ for $1 \leq j < k \leq n$
DifferentIndividuals(i_1 … i_n)	$S(i_j) \neq S(i_k)$ for $1 \leq j < k \leq n$
Individual([i] annotation(…) … annotation(…) type(c_1) … type(c_m) pv_1 … pv_n)	EC(Individual([i] type(c_1) … type(c_m) pv_1 … pv_n)) is nonempty

The effect of an imports construct is to import the contents of another OWL ontology into the current ontology. The imported ontology is the one that can be found by accessing the document at the URI that is the argument of the imports construct. The *imports closure* of an OWL ontology is then the result of adding the contents of imported ontologies into the current ontology. If these contents contain further imports constructs, the process is repeated as necessary. A particular ontology is never imported more than once in this process, so loops can be handled.

Annotations have no effect on the semantics of OWL ontologies in the abstract syntax.

An Abstract OWL interpretation, I, is an interpretation of an OWL ontology, O, iff I is an interpretation of each axiom and fact in the imports closure of O.

An Abstract OWL ontology *entails* an OWL axiom or fact if each interpretation of the ontology is also an interpretation of the axiom or fact. An Abstract OWL ontology entails another Abstract OWL ontology if each interpretation of the first ontology is also an interpretation of the second ontology. Note that there is no need to create the imports closure of an ontology—any method that correctly determines the entailment relation is allowed.

From the RDF model theory [*RDF MT*], for V a set of URI references containing the RDF and RDFS vocabulary, an RDFS interpretation over V is a triple $I=<R_I, EXT_I, S_I>$. Here R_I is the domain of discourse or universe, i.e., a set that contains the denotations of URI references. EXT_I is used to give meaning to properties and is a mapping from R_I to sets of pairs over $R_I \times (R_I \cup LV)$. Finally, S_I is a mapping from V to R_I that takes a URI reference to its denotation. $CEXT_I$ is then defined as $CEXT_I(c) = \{ x \in R_I \mid <x,c> \in EXT_I(S_I(\text{rdf:type})) \}$. RDFS interpretations must meet several conditions, as detailed in the RDFS model theory. For example, S_I(rdfs:subClassOf) must be a transitive relation.

An OWL interpretation, $I=<R_I, EXT_I, S_I>$, over a vocabulary V, where V includes VRDFS, rdfs:Literal, VOWL, owl:Thing, and owl:Nothing, is an RDFS interpretation over V that satisfies the following conditions:

Relationships between OWL classes

If E is	then $CEXT_I(S_I(E))=$	with
owl:Thing	IOT	$IOT \subseteq R_I$
owl:Nothing	{}	
rdfs:Literal	LV	
owl:Class	IOC	$IOC \subseteq CEXT_I(S_I(\text{rdfs:Class}))$
owl:Restriction	IOR	$IOR \subseteq IOC$
owl:Datatype	IDC	$IDC \subseteq CEXT_I(S_I(\text{rdfs:Class}))$
owl:Property	IOP	$IOP \subseteq CEXT_I(S_I(\text{rdf:Property}))$
owl:ObjectProperty	IOOP	$IOOP \subseteq IOP$
owl:DataTypeProperty	IODP	$IODP \subseteq IOP$
rdf:List	IL	$IL \subseteq R_I$

Membership in OWL classes

If E is	then $S_I(E)$
owl:Thing	IOC
owl:Nothing	IOC
rdfs:Literal	IDC
a datatype of D	IDC
rdf:nil	IL

Characteristics of members of OWL classes

If E is	then if $e \in CEXT_I(S_I(E))$ then
owl:Class	$CEXT_I(e) \subseteq IOT$
owl:Datatype	$CEXT_I(e) \subseteq LV$
owl:ObjectProperty	$EXT_I(e) \subseteq IOT \times IOT$
owl:DatatypeProperty	$EXT_I(e) \subseteq IOT \times LV$

The next constraints are IFF, which may be harder to deal with in OWL/DL, as they extend the various categories of properties to all of owl:Property. However, in OWL/DL ontologies, you can neither state that an owl:DatatypeProperty is inverse functional nor ask whether it is, so there should be not adverse consequences.

If E is	then $c \in CEXT_I(S_I(E))$ iff $c \in$ IOP and
owl:SymmetricProperty	$<x,y> \in EXT_I(c) \rightarrow <y, x> \in EXT_I(c)$
owl:FunctionalProperty	$<x,y_1>$ and $<x,y_2> \in EXT_I(c) \rightarrow y_1 = y_2$
owl:InverseFunctionalProperty	$<x_1,y> \in EXT_I(c) \cap <x_2,y> \in EXT_I(c) \rightarrow x_1 = x_2$
owl:TransitiveProperty	$<x,y> \in EXT_I(c) \cap <y,z> \in EXT_I(c) \rightarrow <x,z> \in EXT_I(c)$

RDFS domains and ranges are strengthened to if-and-only-if over the OWL universe

If E is	then for	$<x,y> \in CEXT_I(S_I(E))$ iff
rdfs:domain	$x \in$ IOP, $y \in$ IOC	$<z,w> \in EXT_I(x) \rightarrow z \in CEXT_I(y)$
rdfs:range	$x \in$ IOP, $y \in$ IOCÎIDC	$<w,z> \in EXT_I(x) \rightarrow z \in CEXT_I(y)$

Some OWL properties have iff characterizations

If E is	**then $\langle x,y\rangle \in EXT_I(S_I(E))$ iff**
owl:sameClassAs	$x,y \in IOC \wedge CEXT_I(x) = CEXT_I(y)$
owl:disjointWith	$x,y \in IOC \wedge CEXT_I(x) \cap CEXT_I(y) = \{\}$
owl:samePropertyAs	$x,y \in IOP \wedge EXT_I(x) = EXT_I(y)$
owl:inverseOf	$x,y \in IOOP \wedge \langle u,v\rangle \in EXT_I(x)$ iff $\langle v,u\rangle \in EXT_I(y)$
owl:sameIndividualAs	$x = y$
owl:sameAs	$x = y$
owl:differentFrom	$x \neq y$

Some OWL properties have only-if characterizations

We will say that l_1 is a sequence of $y_1,\ldots,y_n$ over C iff n=0 and $l_1 = S_I(\text{rdf:nil})$ or n>0 and $l_1 \varepsilon IL$ and $\exists l_2, \ldots, l_n \in IL$ such that $\langle l_1,y_1\rangle \in EXT_I(S_I(\text{rdf:first}))$, $y_1 \in CEXT_I(C)$, $\langle l_1,l_2\rangle \in EXT_I(S_I(\text{rdf:rest}))$, …, $\langle l_n,y_n\rangle \in EXT_I(S_I(\text{rdf:first}))$, $y_n \in CEXT_I(C)$, and $\langle l_n,S_I(\text{rdf:nil})\rangle \in EXT_I(S_I(\text{rdf:rest}))$.

If E is	**then if $\langle x,y\rangle \in EXT_I(S_I(E))$ then**
owl:complementOf	x,y∈ IOC and CEXTI(x)=IOT-CEXTI(y)
If E is	**then if $\langle x,l\rangle \in EXT_I(S_I(E))$ then**
owl:unionOf	x∈ IOC and l is a sequence of $y_1,\ldots y_n$ over IOC and $CEXT_I(x) = CEXT_I(y_1) \cup \ldots \cup CEXT_I(y_n)$
owl:intersectionOf	x∈ IOC and l is a sequence of $y_1,\ldots y_n$ over IOC and $CEXT_I(x) = CEXT_I(y_1) \cap \ldots \cap CEXT_I(y_n)$
owl:oneOf	x∈ $CEXT_I$(rdfs:Class) and l is a sequence of $y_1,\ldots y_n$ over $R_I \cup LV$ and $CEXT_I(x) = \{y_1,\ldots, y_n\}$

If E is	**and**	**then if $\langle x,l\rangle \in EXT_I(S_I(E))$ then**
owl:oneOf	l is a sequence of $y_1,\ldots y_n$ over LV	x∈ IDC and $CEXT_I(x) = \{y_1,\ldots, y_n\}$
owl:oneOf	l is a sequence of $y_1,\ldots y_n$ over IOT	x∈ IOC and $CEXT_I(x) = \{y_1,\ldots, y_n\}$

If	**then x∈ IOR, y∈ IOC, p∈ IOP, and $CEXT_I(x)$ =**
$\langle x,y\rangle \in EXT_I(S_I(\text{owl:allValuesFrom}))) \wedge \langle x,p\rangle \in EXT_I(S_I(\text{owl:onProperty})))$	$\{u \in IOT \mid \langle u,v\rangle \in EXT_I(p) \rightarrow v \in CEXT_I(y)\}$
$\langle x,y\rangle \in EXT_I(S_I(\text{owl:someValuesFrom}))) \wedge \langle x,p\rangle \in EXT_I(S_I(\text{owl:onProperty})))$	$\{u \in IOT \mid \exists \langle u,v\rangle \in EXT_I(p) \wedge v \in CEXT_I(y)\}$
$\langle x,y\rangle \in EXT_I(S_I(\text{owl:hasValue}))) \wedge \langle x,p\rangle \in EXT_I(S_I(\text{owl:onProperty})))$	$\{u \in IOT \mid \langle u, y\rangle \in EXT_I(p)\}$

If	**then x∈ IOR, y∈ LV, y is a nonnegative integer, p∈ IOP, and $CEXT_I(x)$ =**
$\langle x,y\rangle \in EXT_I(S_I(\text{owl:minCardinality}))) \wedge \langle x,p\rangle \in EXT_I(S_I(\text{owl:onProperty})))$	$\{u \in IOT \mid \text{card}(\{v : \langle u,v\rangle \in EXT_I(p)\}) \geq y\}$
$\langle x,y\rangle \in EXT_I(S_I(\text{owl:maxCardinality}))) \wedge \langle x,p\rangle \in EXT_I(S_I(\text{owl:onProperty})))$	$\{u \in IOT \mid \text{card}(\{v : \langle u,v\rangle \in EXT_I(p)\}) \leq y\}$
$\langle x,y\rangle \in EXT_I(S_I(\text{owl:cardinality}))) \wedge \langle x,p\rangle \in EXT_I(S_I(\text{owl:onProperty})))$	$\{u \in IOT \mid \text{card}(\{v : \langle u,v\rangle \in EXT_I(p)\}) = y\}$

R_I contains elements corresponding to all possible OWL descriptions and data ranges

The first three conditions require the existence of the finite sequences that are used in some OWL constructs. The remaining conditions require the existence of the OWL descriptions and data ranges.

If there exists	**then there exists $l_1,\ldots,l_n \in$ IL with**
$x_1, \ldots, x_n \in IOC$	$\langle l_1,x_1\rangle \in EXT_I(S_I(\text{rdf:first}))$, $\langle l_1,l_2\rangle \in EXT_I(S_I(\text{rdf:rest}))$, … $\langle l_n,x_n\rangle \in EXT_I(S_I(\text{rdf:first}))$, $\langle l_n,S_I(\text{rdf:nil})\rangle \in EXT_I(S_I(\text{rdf:rest}))$
$x_1, \ldots, x_n \in IOT \cup LV$	$\langle l_1,x_1\rangle \in EXT_I(S_I(\text{rdf:first}))$, $\langle l_1,l_2\rangle \in EXT_I(S_I(\text{rdf:rest}))$, … $\langle l_n,x_n\rangle \in EXT_I(S_I(\text{rdf:first}))$, $\langle l_n,S_I(\text{rdf:nil})\rangle \in EXT_I(S_I(\text{rdf:rest}))$

If there exists	then there exists y with
l, a sequence of $x_1,\ldots,x_n$ over IOC	$y \in IOC, <y,l> \in EXT_I(S_I(owl:unionOf))$
l, a sequence of $x_1,\ldots,x_n$ over IOC	$y \in IOC, <y,l> \in EXT_I(S_I(owl:intersectionOf))$
l, a sequence of $x_1,\ldots,x_n$ over IOT∪LV	$y \in CEXT_I(S_I(rdfs:Class)), <y,l> \in EXT_I(S_I(owl:oneOf))$
If there exists	**then there exists y ∈ IOC with**
$x \in IOC$	$<y,x> \in EXT_I(S_I(owl:complementOf))$
If there exists	**then there exists y ∈ IOR with**
$x \in IOP \wedge w \in IOC \cup IDC$	$<y,x> \in EXT_I(S_I(owl:onProperty)) \wedge <y,w> \in EXT_I(S_I(owl:allValuesFrom))$
$x \in IOP \wedge w \in IOC \cup IDC$	$<y,x> \in EXT_I(S_I(owl:onProperty)) \wedge <y,w> \in EXT_I(S_I(owl:someValuesFrom))$
$x \in IOP \wedge w \in IOT \cup LV$	$<y,x> \in EXT_I(S_I(owl:onProperty)) \wedge <y,w> \in EXT_I(S_I(owl:hasValue))$
$x \in IOP \wedge w \in LV \wedge w$ is a nonnegative integer	$<y,x> \in EXT_I(S_I(owl:onProperty)) \wedge <y,w> \in EXT_I(S_I(owl:minCardinality))$
$x \in IOP \wedge w \in LV \wedge w$ is a nonnegative integer	$<y,x> \in EXT_I(S_I(owl:onProperty)) \wedge <y,w> \in EXT_I(S_I(owl:maxCardinality))$
$x \in IOP \wedge w \in LV \wedge w$ is a nonnegative integer	$<y,x> \in EXT_I(S_I(owl:onProperty)) \wedge <y,w> \in EXT_I(S_I(owl:cardinality))$

References

1. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large-scale vocabulary test. J Am Med Inform Assoc. 1997; 4(6): 484–500.
2. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a Large Controlled Medical Terminology. JAMIA. 1994; 1(1):35–50.
3. Elkin PL, Chute CG, et al. Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the unified medical language system. JAMIA, Symp. Suppl., 1997:500–4
4. Rassinoux AM, Miller RA, Baud R, Scherrer JR. Modeling just the important and relevant concepts in medicine for medical language understanding: a survey of the issues. In Chute C, editor. Proceedings of the IMIA WG-6, Jacksonville; 1997. p 53–68.
5. Rector AL, Nowlan WA. The Galen project. Comput Methods Programs Biomed. 1994;45(1–2):75–8.
6. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS, for the Canon Group. Toward a medical-concept representation language. JAMIA. 1994;1:207–17.
7. Wagner JC, Rogers JE, Baud RH, Scherrer JR. Natural language generation of surgical procedures. Int J Med Inform. 1999;53(2–3):175–92.
8. Wagner JC, Rogers JE, Baud RH, Scherrer JR. Natural language generation of surgical procedures. Medinfo. 1998;9(Pt 1):591–5.
9. Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux RAM, Abrial V, Clavel L, Idir H. Galen: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. Stud Health Technol Inform. 1999;68:901–5.
10. Wroe CJ, Cimino JJ, Rector AL. Integrating existing drug formulation terminologies into an HL7 standard classification using OpenGALEN. Proceedings/ AMIA Annu Symp:766–70; 2001.
11. Rector AL. Clinical terminology: why is it so hard? Methods Inf Med. 1999 Dec;38(4–5):239–52.
12. Rector AL. The interface between information, terminology, and inference models. Medinfo. 2001;10(Pt 1): 246–50.
13. Elkin PL, Cimino JJ, Lowe HJ, Aronow DB, Payne TH, Pincetl PS, Barnett GO. Mapping to MeSH. Presented to, and Published in the IEEE proceedings of the 12th annual symposium on computers and medical care. 01954210/88/0000/0185$01.00 © 1988 SCAMC, Inc.
14. Cimino JJ, Mallon LJ, Barnett GO. Automated extraction of medical knowledge from medline citations. Presented to, and Published in the IEEE proceedings of the 12th annual symposium on computers and medical care, Washington, DC, 1988.
15. Baud R, Rassinoux A, Scherrer J. Natural language processing and semantically representation of medical texts. Meth of Inf Med. 1992;31(2):117–25.
16. Baud R, Lovis C, Rassinoux AM, Scherrer JR. Alternate ways of knowledge collection, indexing and robust language retrieval. In Chute C, editor. Proceedings of the IMIA WG-6, 1997. p 81–93.
17. http://www.w3.org/TR/owl-semantics/#1
18. Elkin PL, Ruggieri AP, Brown SH, Buntrock J, Bauer BA, Wahner-Roedler D, Litin S, Beinborn J, Bailey KR, Bergstrom L. A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM. JAMIA Suppl. 2001: 159–63
19. Elkin PL, Tuttle M, Keck K, Campbell K, Atkin G, Chute CG. The role of compositionality in standardized problem list generation. In Cesnik B, editor. Proceedings of MedInfo, Amsterdam; 1998.

Interface Terminologies

6

S. Trent Rosenbloom

Note to the reader: This chapter assembles and digests content from several articles published in the biomedical literature. The complete articles are listed in the reading list at the end of this chapter. The two key articles are:

- Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13:277–88.
- Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. A model for evaluating interface terminologies. J Am Med Inform Assoc 2008;15:65–76.

Electronic health record (EHR) systems continue to gain traction across the United States. A major promise of EHR systems is their capacity to help healthcare providers capture structured clinical information directly during patient care. Structured information can be defined as that information which is represented in a standard's complaint fashion and which is designed to meet a specific purpose. Researchers have noted that a common obstacle to EHR system adoption is the difficulty in creating tools that assist healthcare providers in documenting structured clinical information. This is due in large part to the fact that clinical information is often complex and unpredictable and that healthcare providers typically prefer to document clinical care using narrative text natural language. Specialized tools called "interface terminologies" are designed to address these obstacles. Interface terminologies bring together commonly used human-friendly phrases (called terms) and associated modifiers to improve the acceptability and efficiency of using structured clinical documentation tools.

Interface terminologies support healthcare providers documenting patient-related information into computer programs such as computer-based documentation and decision support systems [1–5]. Such interface terminologies "interface" between clinicians' own unfettered, colloquial conceptualizations of patient descriptors and the more structured, coded internal data elements used by specific healthcare application programs. Interface terminologies generally embody a rich set of flexible phrases displayed in the graphical or text interfaces of specific computer programs. Among their many applications, interface terminologies have been used for problem list entry, clinical documentation in electronic health record systems, text generation, care provider order entry with decision support, and diagnostic expert programs.

This chapter will review the goals of terminologies in general, will discuss the goals and important attributes for interface terminologies, and will address methods for evaluating interface terminologies.

S.T. Rosenbloom, M.D., MPH, FACMI
Department of Biomedical Informatics,
Vanderbilt University Medical Center,
Nashville, TN, USA
e-mail: trent.rosenbloom@vanderbilt.edu

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_6,

Clinical Terminologies

Terminologies are a type of software made up of compilations of words or phrases, collectively called terms, typically assembled together in a systematic fashion with the goal of representing the entities and events (i.e., the conceptual information) that makes up a particular knowledge domain. For example, a given terminology representing the domain of clinical medicine may include the terms "myocardial infarction" or "heart attack" to represent the event "ischemic injury and necrosis of heart muscle cells resulting from absent or diminished blood flow in a coronary artery." A clinician who is taking care of a patient who has recent onset of crushing chest pain may consider a diagnosis of ischemic heart muscle injury. The clinician would communicate the diagnosis using either of terms "myocardial infarction" or "heart attack." Terminologies often organize their concepts along a hierarchical representation, made up of linkages among the concepts. One commonly used linkage is the hierarchical "Is-a-type-of" relationship, such as might be observed between "myocardial infarction" and "heart disease" in a hypothetical terminology [6–8].

Over the past few centuries, numerous clinical terminologies have been developed. However, no single terminology has emerged to serve as a universal standard that can represent a majority of clinical concepts for all uses. While no terminology has been identified as a standard to cover all of health care, numerous individual terminologies may serve specific needs. In this spirit, the United States National Committee on Vital and Health Statistics (NCVHS) and the United States government's multiagency consolidated health informatics (CHI) council recommended in 2003 that healthcare providers leverage a set of existing terminologies as standards for focused aspects of clinical knowledge and information. The NCVHS asserted that the terminologies "(1) are required to adequately cover the domain of patient medical record information and (2) meet essential technical criteria to serve as *reference terminologies*." Some of the NCVHS-recommended terminologies include SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) to be used for "the exchange, aggregation, and analysis of patient medical information," LOINC (Logical Observation Identifiers Names and Codes) to be used when representing individual laboratory and other diagnostic tests, and several specific drug terminologies (e.g., RxNorm and the National Drug File Reference Terminologies [NDF-RT]) when representing medications, their clinical mechanisms of actions, and physiologic effects. Given the large number of available terminologies, the United States National Library of Medicine (NLM) has for the past two decades worked to create and maintain the Unified Medical Language System (UMLS). The UMLS is designed to assemble multiple terminologies in a thesaurus that aligns their concepts by meaning [9].

Cimino's "Desiderata for Controlled Terminologies" [10] defined essential attributes of a "sharable, multipurpose" terminology. Desiderata emphasized the importance of concept orientation during terminology construction, which involves using concepts as "basic building blocks" rather than words, terms, or phrases. Desiderata also emphasized the importance that a terminology covers its target domain's concepts completely and at multiple levels of detail. Desiderata defined a number of additional desired attributes for formal terminologies. Other terminology functional standards echoed the Desiderata, such as those in "A Framework for Comprehensive Health Terminology Systems in the United States" [11] and the International Standards Organization's technical specifications [12, 13].

Interface Terminologies

Large clinical terminologies that have precise terms to represent concepts may not always be usable by healthcare providers. However, a key desired attribute for a terminology is to have adequate domain coverage to be useful. Because there is a need to balance domain coverage with easy usability, various stakeholders and investigators have suggested that terminology developers build terminologies specifically for specific needs. For example, terminologies can be created

to meet the specific needs of administrative processes such as billing, formally representing concepts and their interrelationships in support of research and data exchange, and promoting efficient recording of common clinical findings and events into problem lists and progress notes.

In 1994, Campbell described (but did not name) interface terminologies as terminologies designed to support efficient structured clinical documentation using computerized note capture tools [4]. Campbell indicated that the way to approach creating such a terminology was to focus on modeling concepts commonly used by healthcare providers. Such terminologies designed to support efficient use by healthcare providers entering clinical information directly were first called "interface terminology" by Kent Spackman in 1997. This name referred to their use in support of data entry in a user interface designed for clinical documentation. Other investigators have also called these "colloquial terminologies," [2, 14] "application terminology," [3] and "entry terminology." [1] Combining these references, an interface terminology can be defined as "a systematic collection of clinically oriented phrases (terms), whose purpose is to support clinicians' entry of patient information into computer programs, such as clinical note capture and decision support tools." [5]

As above, Cimino's desiderata described a number of desired attributes that should be present in clinical terminologies. Interface terminologies need additional attributes to improve the expressivity and usability they support. Table 6.1 lists Cimino's desiderata alongside additional desiderata pertinent to interface terminologies. Interface terminology usability, as with usability in general, indicates that users can efficiently, easily, and with satisfaction, accomplish the tasks they intend while using it. In the case of interface terminologies, usability relates to the ability for the healthcare provider to document patient care when using a clinical documentation system that overlies the terminology. There are six desiderata relevant to interface terminology usability. These are (1) completeness of synonym coverage; (2) a balance between precoordination and postcoordination; (3) inclusion of adequate and relevant assertional medical knowledge, as defined below;

Table 6.1 Four ways to compose "appendicitis" using SNOMED RT, as initially described by the CANON Group [29]

D5-46210 01 Acute appendicitis, NOS	G-A231 01 Acute D5-46100 01 Appendicitis, NOS
M-41000 01 Acute inflammation, NOS	G-A231 01 Acute
G-CO06 01 In	M-40000 01 Inflammation, NOS
T-59200 01 Appendix, NOS	G-CO06 01 In T-59200 01 Appendix, NOS

(4) mapping to terminologies having more formal concept representations; (5) support for human-readable output; and (6) being independent from the computer program that uses it. These desiderata are discussed in detail as follows.

Synonymy in an Interface Terminology

Different terms that make up alternative representations for the concepts in a terminology are called synonyms. For example, the phrase "MRI" is a synonym (and abbreviation) for "magnetic resonance imaging scan." Synonyms can help healthcare providers using terminologies to find terms that are familiar to them and therefore the underlying concepts. An adequate richness of synonyms in an interface terminology can increase its usability. A variety of different types of synonyms exist, including alternate terms (e.g., "myocardial infarction" for "heart attack"), acronyms (e.g., "MI" for "myocardial infarction"), definitional phrases (e.g., "ischemic injury and necrosis of heart muscle cells resulting from absent or diminished blood flow in a coronary artery"), and eponyms (e.g., "Levine sign" for "a clenched fist held over the chest indicating ischemic cardiac chest pain") [15]. Interface terminologies should embody the richness present in the colloquial phrases of medical discourse. Interface terminologies having a rich synonymy can support a nuanced approach to clinical documentation, with which healthcare providers can express themselves fluidly. A downside to a rich synonymy is that synonyms may also increase the chances that a given term may be used to represent more than one concept (e.g., "cold" for "a low temperature" and for

"upper respiratory tract viral infection"). In this situation, excess synonymy may increase a terminology's ambiguity.

Two different attributes can be used to indicate how well a terminology's synonymy represents a clinical domain. These attributes are called accuracy and expressivity. In a terminology, *accuracy* indicates how closely a term's meaning corresponds with the underlying concept it represents. For example, a healthcare provider would likely agree that "heart attack" is an accurate synonym for representing the concept "myocardial infarction," while the synonym "acute myocardial infarction" would be less accurate because it is more specific. Unlike accuracy, synonym *expressivity* reflects how well a term's *semantic character* matches the words in the phrase it is meant to represent rather than the underlying meaning. Semantic character can be defined as the narrative flavor, implicit clinical urgency, and specificity of meaning conveyed by the words and the word order in a given clinical phrase. The presence of expressivity in an interface terminology reduces the time healthcare providers require when entering their own "natural language" terms into a structured documentation system. The adequacy of a terminology's expressivity is judgmental in nature in that different users may reasonably disagree about whether the terminology's terms adequately convey semantic character or nuance.

The nature of the attributes accuracy and expressivity can be demonstrated in the following example. In an actual case in our healthcare center, a patient once described to his cardiologist that he had a "feathery discomfort occurring across the chest." It is likely that a clinical terminology would include a concept represented by a term such as "chest discomfort," which might even have associated modifiers "soft" and "anterior chest wall." However, it is unlikely that the terminology would include a modifier "feathery." A healthcare provider might agree that the two modifiers "noncrushing" and "feathery" are reasonable synonyms when detailing chest discomfort. However, because each uses different terms and may evoke to a healthcare provider different semantic nuances. That is, a concept assembling the components from the terminology to compose "noncrushing chest discomfort" to represent the patient's statement of "feathery chest discomfort" is accurate because the two phrases reasonably have the same clinical meaning. However, to a patient or healthcare provider, the first phrase does not fully express the character of the second.

Balancing Precoordination and Postcoordination in an Interface Terminology

As above, a major goal for terminologies is to cover their intended domain completely and comprehensively with concepts and relationships among them. Terminology developers can create large terminologies to cover knowledge domains through one of two general approaches. In one approach, developers enumerate (i.e., "precoordinate") all possible complex concepts a priori. This approach essentially creates a listing of all the complex concepts that can be expressed within the terminology. This approach may increase the chances that a terminology user will find a desired concept. There are two primary disadvantages to precoordination in a terminology. First, a large, extensive would make the terminology so large that a user might have difficulty searching through it. For example, a precoordinated pharmacy formulary terminology might list all medications in all combinations of dose form and strength. So the terminology would have separate entries for "amoxicillin 875 mg tablet," "amoxicillin 500 mg tablet," and "amoxicillin 400 mg/5 mL suspension," among others. A pharmacist using this terminology would have to search through all combinations of several dozen generic and brand names with three different dosage strengths each – potentially as many as 50–100 items to choose from when encoding an antibiotic as amoxicillin. Second, precoordinated terminologies may be relatively inflexible in situations where they do not contain concepts that a user may need for a given task.

An alternative to precoordination is an approach that allows users to compose complex concepts by assembling numerous relatively

general concepts and modifiers as needed. This approach, called postcoordination, can increase a terminology's flexibility for representing a wide range of concepts. In a terminology permitting postcoordination, existing concepts typically are modeled at a fairly general level of detail. Additional complexity is modeled when appropriate for a given task. In this way, the overall terminology size can be kept relatively small, while the number of complex concepts that can potentially be constructed can be fairly large. There are three primary disadvantages to postcoordination in a terminology. First, because different users might compose similar concepts in different ways, postcoordination may decrease the chance that they apply the terminology consistently. Table 6.2 provides an example of several different ways a user might appropriately compose the simple concept, "appendicitis." Second, using a postcoordinated terminology, users can create nonsensical complex concepts when composing them from concepts and modifiers across multiple axes. Third, the process of creating complex postcoordinated concepts during the process of healthcare delivery may be time-consuming and inefficient.

In an interface terminology, using precoordination during terminology design may complement allowing users to postcoordinate concepts as needed. Bringing these two approaches together can optimize a terminology's flexibility, ease of use, and overall coverage. This balance, called "compositional balance," makes concept selection tasks more efficient by reducing the effort required to assemble relatively complex concepts from more general ones and the time needed to browse or search through long lists of precoordinated concepts.

The degree of compositionality in a terminology can be quantified. Campbell has described a method for measuring the "degrees of freedom" intrinsic to any concept in a terminology [16].

Table 6.2 Relative importance of terminology attributes to an interface terminology and to a clinical terminology

Terminology attribute	Clinical terminology	Interface terminology
Statement of purpose, scope, and comprehensiveness	√	√
Complete coverage of domain-specific content	√	√
Use of concepts rather than terms, phrases, and words (concept orientation)	√	
Concepts do not change with time, view, or use (concept consistency)	√	√
Concepts must evolve with change in knowledge	√	√
Concepts identified through nonsense identifiers (context-free identifier)	√	√
Representation of concept context consistently from multiple hierarchies	√	
Concepts have single explicit formal definitions	√	√
Support for multiple levels of concept detail	√	√
Absence of, or methods to identify, duplication, ambiguity, and synonymy	√	
Synonyms uniquely identified and appropriately mapped to relevant concepts	√	√
Support for compositionality to create concepts at multiple levels of detail	√	√
Language independence	√	
Integration with other terminologies	√	
Mapping to administrative terminologies	√	
Complete coverage by domain-specific terms and synonyms		√
Presence of assertional knowledge		√
Presence of optimal compositional balance		√

Reproduced from [5]

The degrees of freedom measure provides a numerical assessment of the complexity of a term. Degrees of freedom are calculated by counting the number of general concepts contained in a complex precoordinated concept. For example, the concept "severe chest pain" can be said to contain three atomic concepts and modifiers "severe," "chest," and "pain." Degrees of freedom can provide a quantitative representation of compositional balance. That is, measuring the degrees of freedom in complex concepts can also expose the number of concepts in a terminology that may be precoordinated and the average number of general concepts used to compose interface concepts across the terminology. There may exist a level of compositional balance that maximizes usability, which may vary by the interface terminology's specific intended use and clinical domain (Fig. 6.1).

Assertional Knowledge in an Interface Terminology

By definition, terminologies represent the knowledge that fills a specific domain, including the concepts and their interrelationships. They can represent two types of knowledge about relationships. One type is called definitional knowledge (which is also called "terminological knowledge" and "contextual knowledge"). Definitional knowledge specifies the purely structural

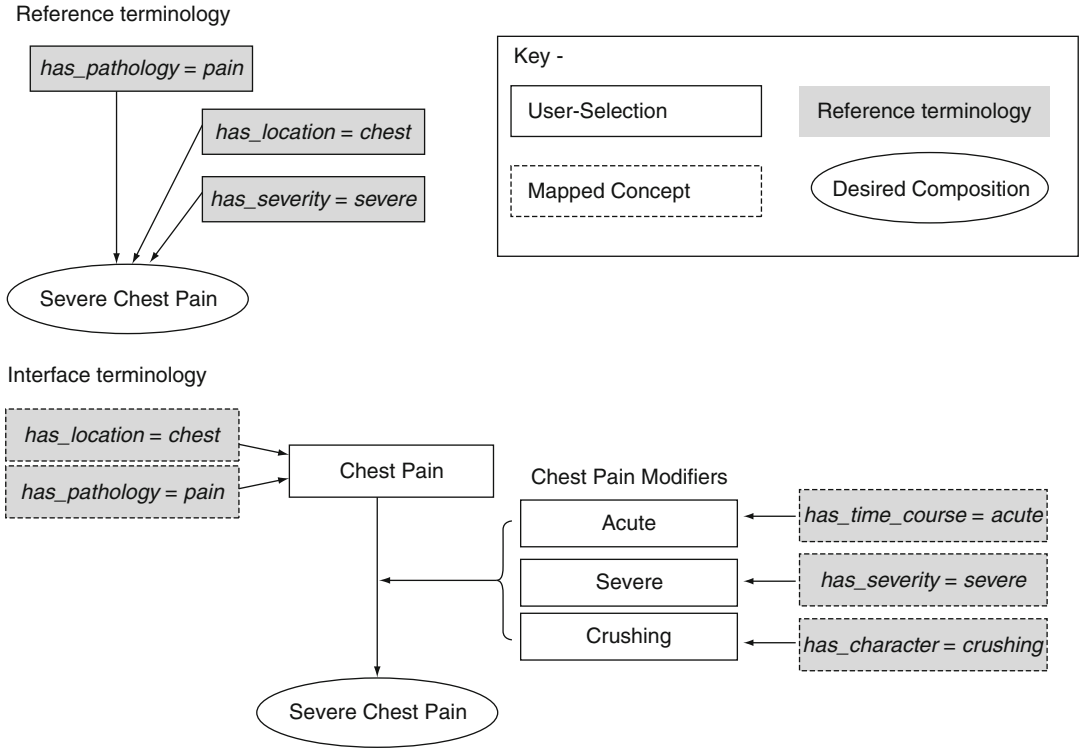


Fig. 6.1 Two approaches to composing the concept "severe chest pain." On the top, a user select concepts and modifiers directly from a reference terminology permitting postcoordination, using description logic to combine unrelated atomic concepts sequentially, starting with "pain" then adding the location modifier "chest" and the severity modifier "severe." On the bottom, the user can combine the precoordinated concept "chest pain" in an interface terminology with the formally linked modifier "severe" from the list of chest pain modifiers. All concepts and modifiers in the interface terminology are mapped to formal representations in an external reference terminology. Both approaches allow the user to compose a meaningful concept having a formal representation (Reproduced from [5])

relationships among concepts. That is, definitional knowledge specifies how concepts' existence relates to other concepts. For example, in the reference terminology SNOMED CT, the concept "chest pain" is formally defined by three definitional relationships: (1) it has an is-a relationship with the concept "finding of region of thorax"; (2) it has an is-a relationship with the concept "pain of truncal structure"; and (3) it has a has-finding-site relationship with the concept "thoracic structure." However, definitional knowledge does not model how concepts that may influence or relate to each other in a clinical setting, such as that "chest pain" may be caused by "myocardial infarction." This latter type of knowledge is called definitional knowledge.

Assertional knowledge is information that provides nuance and context to a concept, but does not specifically define it [17]. Interface terminologies should include assertional knowledge–based relationships in addition to definitional knowledge. Assertional knowledge can model clinically oriented relationships among concepts and modifiers, for example, relationships describing whether concepts are present or absent during certain clinical conditions (e.g., whether chest pain is normally present in a patient experiencing an acute myocardial infarction). Assertional knowledge can also represent relationships between clinical concepts and particular patient populations (e.g., that pregnancy is not present in men or in women who have had a hysterectomy or who are postmenopausal). Assertional knowledge–based relationships can also clarify whether two potential synonyms truly have the same meaning (i.e., are accurate representations of the same concept, as defined above). For example, the terms "thorax pain" and "chest pain" may have the same formal definition and represent the same underlying concept. However, the term "thorax pain" may imply to a healthcare provider that the patient's pain is present in the chest wall (i.e., in the ribs or sternum), while the term "chest pain" might imply that it is more internally located, such as might occur from a cardiac or pulmonary disease. Assertional knowledge–based relationships may include attributes that would distinguish "thorax pain" from "chest pain," such as by including relevant synonyms, associated diagnoses, common symptoms, usual modifiers, and describing prevalence in a given patient population.

In an interface terminology, assertional knowledge is commonly modeled into lists of concepts, synonyms, and to modifiers that are commonly associated with a given concept or term. Assertional knowledge may be more relevant to clinical users than definitional knowledge and may improve their ability to use the terminology efficiently. For example, in the interface terminology CHISL (Categorical Health Information Structured Lexicon, used for clinical documentation at the Vanderbilt University Medical Center and described further later in this chapter) [18], the concept for "chest pain" include links to the normal status modifier "absent" (e.g., "chest pain" is normally absent in a healthy population), to a list of severity modifiers (e.g., "mild," "moderate," and "severe"), and to lists of common associated concepts (e.g., nausea, depression, diaphoresis, anxiety).

The two major goals of including assertional knowledge in an interface terminology are (1) to enhance usability and (2) to improve documentation quality when using a terminology. Assertional knowledge–based links enhance usability by bringing together concepts and modifiers a user is likely to consider and document together. By linking together clinically related content, an interface terminology makes it easy for a user to address each concept without needing to search the entire terminology for each, decreasing the amount of work that users require to find or compose the concepts needed when documenting [19–22]. In the example of "anterior chest pain," a user working with a reference terminology that requires postcoordination would potentially need to compose it from two distinct concepts, one for "anterior chest" and one for "chest pain," assembling them through definitional relationships (i.e., "anterior chest pain" *is-a* "chest pain," *has-finding-site* "anterior chest"). With this approach, the healthcare provider documenting the clinical case might have to take additional steps to add more concepts or modifiers (e.g., "mild"). However, a healthcare provider using an interface terminology that includes assertional knowledge–based

linkages when documenting the concept chest pain is likely to find a precoordinated concept "chest pain" with linked lists of allowable qualifiers, including chest location (e.g., "anterior chest") and severities (e.g., "mild," "moderate," "severe"). In addition to improving documentation efficiency, the assertional knowledge–defined linkages sanction how a user can use and modify concepts, reducing the chance that they will create nonsensical compositions or inadvertently use the incorrect concept when documenting a clinical case.

Mapping Interface Terminologies

A model of the relationships among concepts (whether definitional or assertional) is important for terminologies providing a complete model of the knowledge domain they represent. An explicit and comprehensive model of the relationships among concepts in a terminology provides a structured representation of its knowledge domain. With this representation, a terminology is more easily used for automated data storage, management, and analysis. For example, and as above, the relationship that defines the concept "anterior chest pain" as a more specific version of the concept "chest pain" can be described by the "is-a" and "has-anatomic-location" definitional relationships (i.e., "anterior chest pain" is-a [type of] "chest pain," has-anatomic-location "anterior chest"). Such definitional relationships are called of description logics. The goal of description logics is formally to model and specify the relationships that exist among concepts and modifiers in a terminology. The goal of including description logics in a terminology is to support tasks related to data and knowledge manipulation, including algorithmic data storage, inferencing, subsumption, classification, management, and analysis.

Interface terminologies are typically created to support clinical documentation by busy healthcare providers who may not value data and knowledge manipulation. As a result, interface terminologies are typically optimized to enable human interaction with structured concepts rather than to provide definitional relationships among the concepts. As above, interface terminologies may benefit from including linkages among concepts and modifiers based on assertional knowledge–based relationships. The goal of these assertional knowledge–based relationships is to improve data acquisition efficiency and workflow. Interface terminology users, by contrast, may not directly benefit from formal description logic-defined relationships among concepts. An alternative to embedding description logic-based definitional knowledge linkages directly in interface terminologies is to link the concepts in an interface terminology to equivalent concepts in reference terminologies that include definitional knowledge. Using this method, the definitional knowledge–based relationships are implied from mapped reference terminologies. As a result, appropriately mapped interface terminologies may not require a formal or complete model of the interrelationships among concepts (e.g., the subset/superset relationship, "anterior chest pain" *is-a* "chest pain").

Support for Human Readability

Because the major goal of interface terminologies is to optimize the human-terminology interface, many include methods to improve the efficiency and clarity of data review by healthcare providers. In particular, interface terminologies can be used to help healthcare providers to access, read, and understand clinical data that have previously been encoded by or represented with a terminology. Interface terminologies designed to support human readability may use a number of different strategies to accomplish this goal. The simplest approach is to use relatively colloquial synonyms so that a clinical application's internally encoded data can be displayed using common words or phrases. A more complex and nuanced approach is for the interface terminology to take advantage of programmatic tags encoding grammatical content and rules to support natural language generation. One approach used by some interface terminologies

to using tagged terminologies is called an augmented transition network (ATN). Terminologies generate natural language using ATNs by including for each concept and modifier tags that specify their preferred colloquial term and the grammatical part of speech it is most likely to take. For example, in the interface terminology CHISL, the concept "chest pain" is tagged as a noun, and the modifiers "anterior" and "dull" as adjectives. When a user selects the interface terms "chest pain," "anterior," "dull," and "present," the application using CHISL evokes an ATN that generates the sentence, "anterior dull chest pain is present." The use by interface terminologies of ATN tagging to support natural language generation is common. It was used, for example, in the mid-1970s by Shortliffe's MYCIN's documentation system [23], in the early 1980s by Miller in the "Attending" anesthesia plan critiquing system [24], and in the 1990s by Poon [25].

Application Independence

Classically, interface terminologies can be incorporated into electronic health record systems or clinical documentation systems through one of two approaches. In the first way, the interface terminology content can be created as an integrated component of the system, with the selectable concepts and modifiers inextricably linked to the user interface elements. Through this approach, the user interface programming directly defines and displays selectable terminology components through the items included in menus of drop-down boxes, buttons, list boxes, and so forth. While these selectable items are interface terminology concepts and modifiers by definition, they cannot easily be separated from the system that uses them and therefore are unlikely to constitute a standalone interface terminology. Interface terminologies installed in this way cannot be modified or updated without making direct changes to the systems that use them. For example, an integrated interface terminology may model the Bright Futures [26, 27] standard for typical pediatric developmental milestones in a pediatric clinical documentation system. If the Bright Futures model changes, as it often does, software developers would need to make actual programmatic modifications to the clinical documentation software (which would likely first need to be prioritized, then be bundled with other software modifications and quality testing, and finally released on an infrequent cycle; the current industry standard is approximately 12–18 months from change request to release).

In a second approach, the interface terminology and the electronic health record system or computer-based documentation system exist as separate application components that share a common data model. With this approach, terminology content development and evolution can be distinct from the software iteration cycles, teams, and priority queues. This approach allows relatively easy integration of external standard terminologies and architectures and may promote data reuse and mapping after it has been captured from the healthcare provider. For example, this approach might permit a clinical problem list documentation system to allow healthcare providers to encode patient problems directly into SNOMED CT. Such a system would be able to take advantage of periodic SNOMED CT updates with little additional local work. There are two major disadvantages to this approach. First, this type of independence allows both the user interface and the interface terminology to influence usability such that they may work against each other. Second, using external standard terminologies may reduce how well they support local functional needs, such as text generation.

Example Interface Terminology: MEDCIN

Among the many interface terminologies available, one of the better-known is called MEDCIN. The MEDCIN terminology has been codeveloped with a clinical documentation system since 1978, when Peter Goltra first developed it as a database of precoordinated clinical findings. The MEDCIN terminology currently contains over 215,000

concepts represented by over 600,000 terms. The structured clinical documentation system overlying the MEDCIN terminology has been implemented in numerous commercial electronic health record systems and by the Department of Defense. The terminology includes concepts from clinical histories, physical examination, tests, diagnoses, and therapies to enable coding of complete patient encounters. The MEDCIN developers primarily use precoordination to support what they call "clinically precise phrasing" that reduces the risk of nonsensical and inefficient compositions. MEDCIN also integrates assertional knowledge–based linkages to support the display of concepts that are clinically relevant to one concept that a user is actively working with. The concepts in the MEDCIN terminology are independent from (i.e., can be extracted and developed separately from) the overlying documentation system and have also been tagged with attributes that support natural language generation from structured clinical documentation. The MEDCIN terminology has been linked to other terminologies, including CPT, ICD-9, ICD-10, DSM-IV, and parts of SNOMED CT.

Example Interface Terminology: CHISL

At the Vanderbilt University Medical Center, developers have created an interface terminology to support structured clinical documentation using several different computer-based documentation systems. This terminology is called the Categorical Health Information Structured Lexicon (CHISL). Developers initially developed CHISL by modifying the terminologies initially designed to support the Internist and the Quick Medical Reference (QMR) diagnostic expert systems, previously developed and described by Miller and Masarie [22, 28]. The CHISL interface terminology has been under development and in use since 1999 and currently contains concepts and modifiers commonly documented in the history and physical examination sections in clinical notes across numerous subspecialties. To achieve a degree of compositional balance, CHISL concepts are generally partially precoordinated but allow further postcoordination using modifiers from lists defined by assertional knowledge–based links. Concepts in CHISL also include several synonyms so that a healthcare provider can represent it in a note according to his or her preference and tags to support natural language generation through an ATN that respects the user-selected synonyms. All concepts, linkages to synonyms and relationships are encoded in the CHISL terminology, and CHISL is independent of the various documentation systems that use it. CHISL has been used in general internal medicine, cardiology, emergency room triage, neurology, and cardiothoracic surgery to document inpatient, outpatient, and postoperative care since 2000. There are currently over 3,800 CHISL concepts, and the terminology is used to generate an average of 200 notes per day through two primary clinical documentation systems.

Discussion

A key part of healthcare providers' daily work involves documenting the details of their interactions with patients. The notes that they generate are rich in clinical information covering all aspects of the patients' health conditions and care delivery. Computer-based documentation tools that encourage structured documentation using a predefined terminology can be difficult to use in the face of the complexity and unpredictability of clinical care. When healthcare providers document care, their workflow typically involves searching for and selecting the best concept to represent clinical findings, indicating its status (e.g., whether it is absent or present), and modifying it with relevant concepts or modifiers. Any documentation system that requires users manually to search an entire terminology or to browse through extensive lists of concepts is likely to be too inefficient for busy healthcare providers while delivering patient care. There are a number of factors that can improve a terminology's usability for clinical documentation, including the presence of a rich and accurate

synonymy, the use of assertional knowledge to define relationships among clinically related concepts and modifiers, and achieving the right balance between including precoordinated concepts and permitting postcoordination. The interface terminology attributes proposed in this chapter directly address system usability and can reduce the terminology's reliance on a given implementation environment.

Questions

1. How would you define interface terminologies?
2. What are the six key attributes of an interface terminology that may not be necessary for reference terminologies?
3. What are the four different types of synonyms? What are examples of each?
4. How would you distinguish synonym accuracy from synonym expressivity?
5. What are the downsides of precoordination and permitting postcoordination in an interface terminology?
6. How are degrees of freedom for a complex concept calculated? How many degrees of freedom are present in the phrase, "severe substernal chest pain radiates to the left jaw" when using SNOMED CT to represent it?
7. What is the difference between definitional knowledge and assertional knowledge? What is an example of each?
8. What are some situations where it would be useful for an interface terminology to support an augmented transition network?
9. If you were to create a tool to help healthcare providers document patients' medical problems in a standardized, structured way, what would you look for in the interface terminology you select?
10. Using the interface terminology you selected to help healthcare providers document medical problems, what are the benefits to implementing it as an application-independent interface terminology? What are the downsides? Which approach do you prefer, and why?

References

1. Chute CG, Elkin PL, Sherertz DD, Tuttle MS. Desiderata for a clinical terminology server. Proc AMIA Symp. 1999:42–6.
2. McDonald FS, Chute CG, Ogren PV, Wahner-Roedler D, Elkin PL. A large-scale evaluation of terminology integration characteristics. Proc AMIA Symp. 1999:864–7.
3. Rose JS, Fisch BJ, Hogan WR, et al. Common medical terminology comes of age, part one: standard language improves healthcare quality. J Healthc Inf Manag. 2001;15:307–18.
4. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Inform Assoc. 1994;1:218–32.
5. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006;13:277–88.
6. Shortliffe EH, Perreault LE, Wiederhold G, Fagan LM. Medical informatics: computer applications in health care. Reading: Addison-Wesley Pub Co.; 1990. p. 37–69.
7. Hammond WE, Stead WW, Straube MJ, Jelovsek FR. Functional characteristics of a computerized medical record. Methods Inf Med. 1980;19:157–62.
8. ISO/TS 17117:2002(E): Health Informatics – controlled health terminology – structure and high-level indicators: Technical Committee ISO/TC 215, Health Informatics; 2002.
9. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc. 1997;4:484–500.
10. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37:394–403.
11. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures. J Am Med Inform Assoc. 1998;5:503–10.
12. ISO 1087–1: Terminology work – vocabulary Part 1: theory and application: Technical Committee TC 37/SC 1; ISO Standards – terminology (principles and coordination); 1996.
13. ISO 1087–2: Terminology work – vocabulary Part 2: computer applications: Technical Committee TC 37/SC 3; ISO Standards – computer applications for terminology; 1996.
14. ASTM 2087:2000: Standard specification for quality indicators for controlled health vocabularies: ASTM Committee E31 on healthcare informatics; 2002.
15. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS:

an exploration of different views of synonymy and quality of editing. J Am Med Inform Assoc. 2005;12(4):486–94.
16. Campbell JR. Semantic features of an enterprise interface terminology for SNOMED RT. Medinfo. 2001;10:82–5.
17. Rector AL, Nowlan WA, Kay S. Conceptual knowledge: the core of medical information systems. In: Lun KC, Deguolet P, Piemme TE, Rienhoff O, editors. Proceedings of the seventh world congress on medical informatics (MEDINFO '92), Geneva; 1992. p. 1420–6.
18. Rosenbloom ST, Brown SH, Froehling D, et al. Using SNOMED CT to represent two interface terminologies. J Am Med Inform Assoc. 2009;16:81–8.
19. Rassinoux AM, Miller RA, Baud RH, Scherrer JR. Modeling just the important and relevant concepts in medicine for medical language understanding: a survey of the issues. In: Proceedings of the IMIA WG6 working conference, Jacksonville; 1997.
20. Horrocks IR. A comparison of two terminological knowledge representation systems [Master's], University of Manchester, Manchester; 1995.
21. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. Artif Intell Med. 1997;9:139–71.
22. Masarie Jr FE, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24:379–400.
23. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res. 1975;8:303–20.
24. Miller PL. Critiquing anesthetic management: the "ATTENDING" computer system. Anesthesiology. 1983;58:362–9.
25. Poon AD, Johnson KB, Fagan LM. Augmented transition networks as a representation for knowledge-based history-taking systems. Proc Annu Symp Comput Appl Med Care. 1992:762–6.
26. Palfrey JS. History of bright futures. Pediatr Ann. 2008;37:135–42.
27. Shaw JS. Practice improvement: child healthcare quality and bright futures. Pediatr Ann. 2008;37:159–64.
28. Miller RA, McNeil MA, Challinor SM, Masarie Jr FE, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project–status report. West J Med. 1986;145:816–22.
29. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Toward a medical-concept representation language. The Canon Group. J Am Med Inform Assoc. 1994;1:207–17.

Springer Terminology Related Standards Development

7

Peter L. Elkin

This chapter will concern itself with just a few of the many standard development organizations (SDOs) that are relevant to healthcare. Also, it will only discuss one or two standards created by an SDO as the volume of work is too large to be discussed in any one textbook. Also, there are many great leaders of informatics who have contributed to terminological standards, many of whom will not be mentioned as we are only providing information that will serve as poignant examples of standards related contributions to the field of healthcare terminologies or terminological representation.

There are many types of standards related to terminologies. These include the individual terminological efforts that have become national or international standards. These are discussed in the "Implementation of Terminology" chapter. There are standards relating to knowledge representation which will be discussed in the "Knowledge Representation and Logic" chapter. In this chapter, we will primarily discuss the health informatics SDOs and some of the related computing standards.

The SDOs and related organizations that will be described in this chapter are:

- HL7
- ASTM E31
- DICOM
- CDISC
- OMG
- OASIS
- IHTSDO
- CEN TC 251
- ISO TC 215
- NCHS

Should the reader be interested in more information regarding a particular SDO or standard, they are referred to that SDO's web site for more information. Each of these SDOs is an open organization and encourages broad participation. Some SDO's do have membership fees associated with participation, and we will try to point that out as we discuss each individual SDO.

P.L. Elkin, M.D., MACP, FACMI
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

Health Level Seven (HL7)

HL7 is an SDO started in the late 1980s and takes its name from the International Standards Organization (ISO) standard networking levels where the seventh level is the application level.

HL7's mission is to create application standards in the healthcare domain. The organization was started in response to a need for messaging standards, and the HL7 V2 messaging standard has been adopted by over 95% of all healthcare institutions as the method for exchanging information between applications in their computer systems. Clem McDonald, M.D., and others who were already heavily involved with health informatics standards was among the thought leaders who created HL7. The organization has

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_7, © Springer-Verlag London 2012

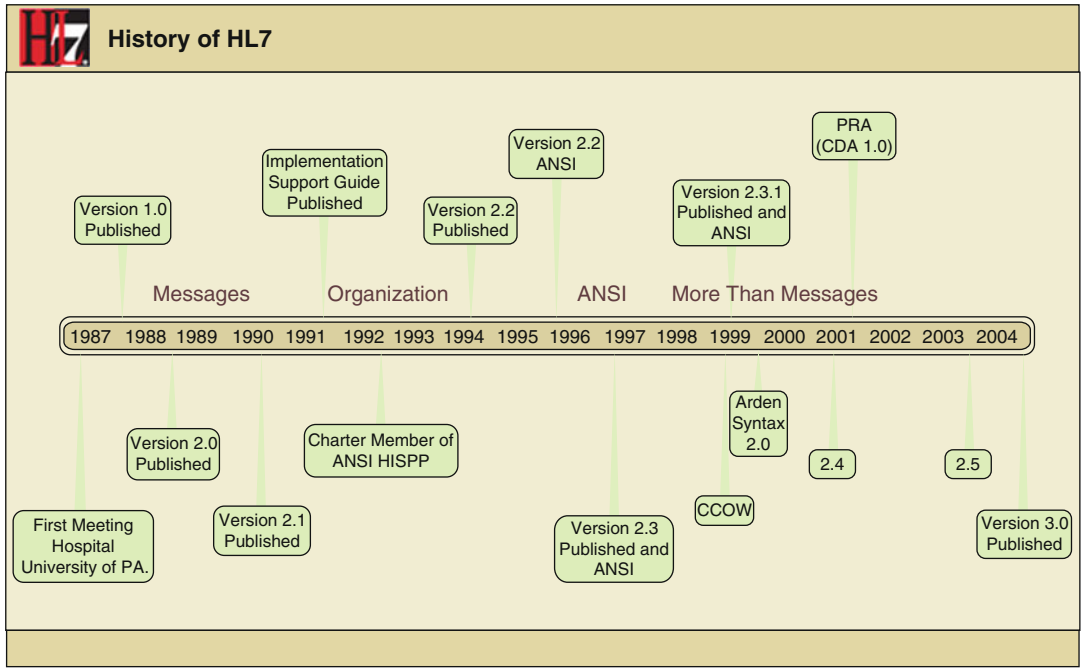


Fig. 7.1 Time line and broad history of standards development at HL7

expanded to model driven messaging standards and standards for decision support and terminological knowledge representation including representation of information for clinical records. Ed Hammond, Ph.D.; Bob Dolin, M.D.; and Stan Huff, M.D., have been among the most consistent and ardent thought leaders behind the representation of clinical data at HL7. HL7 meets at least three times a year with often one meeting outside of the United States and has broad participation by healthcare organizations, industry, and government members (Fig. 7.1) [1].

Version 2.x Messaging Standard

HL7 V2 messages, formally published as "Application Protocol for Electronic Data Exchange in Healthcare Environments," are a set of interoperability specification for transactions created and received by and to computer systems. These specifications describe the transaction interactions by domain and are published as a collection of chapters.

HL7's version 2.x messaging standard is the standard for electronic data exchange in the clinical domain and is the most widely implemented HL7 standard for healthcare in the world. To date, there have been seven releases of the version 2.x standard.

The HL7 V2 standard covers messages that exchange information in the areas of:

- Patient demographics
- Patient charges and accounting
- Patient insurance and guarantor
- Clinical observations
- Encounters including registration, admission, discharge, and transfer
- Orders for clinical service (tests, procedures, pharmacy, dietary, and supplies)
- Observation reporting including test results
- The synchronization of master files between systems
- Medical records document management
- Scheduling of patient appointments and resources
- Patient referrals – Specifically messages for primary care referral
- Patient care and problem-oriented records

This represents most of the clinical and administrative data necessary to run a practice and bill for the services provided. What is missing is the

context that a model-based approach would provide, and therein lies the rationale for HL7 moving to create its version 3 messaging standard.

Version 3 Messaging Standard

HL7 V3 messages are an interoperability specification for transactions that are derived from the HL7 V3 foundation models and vocabulary which provide XML formats for health informatics communications produced and received by computer systems. V3 messages include the concepts in the message wrappers, sequential interactions, and model-based message constructs. These specifications are published by domain as a collection of topics that describe the transactions and their interactions.

The models are governed by the HL7 Reference Information Model with the core classes highlighted in Fig. 7.2.

This model specifies a method of interaction and representation of healthcare knowledge. This is somewhat at odds with the terminological model as there are some types of information that can, at the modelers' discretion, be implemented at the information model or the terminological model, leading to ambiguous representations. For example, negation can be represented within terms or within the information model (e.g., *No Known Drug Allergies*).

V3 Messages include the concepts of message wrappers, sequential interactions, and model-based message content which they call payloads.

HL7 V3 Foundation and Infrastructure

- *HL7 Version 3 Standard: Common Message Element Types, R1*: Common message element types (CMETs) are standardized model fragments intended to be building blocks or components that can be reused by modelers for individual content domains. These components reduce the effort needed to produce a domain-specific design and assure that similar models are implemented across diverse domains, leading to greater consistency of the knowledge representation.
- *HL7 Version 3 Standard: Infrastructure Management, R1*: This document includes information from the Transmission Infrastructure, Control Act Infrastructure, Master File Infrastructure, and the Query Infrastructure domains.
- *HL7 Version 3 Standard: Refinement, Constraint, and Localization to Version 3 Messages, R2*: This document describes the processes describing how HL7 V3 message specifications can be refined, constrained, and extended to support implementation designs, conformance profiles, and realm-specific (locality specific) standards.
- *HL7 Version 3 Standard: Shared Messages, R2*: This document provides information regarding common messages such as acknowledgments that are shared across multiple domains.
- *HL7 Version 3 Standard: Transport Specification, MLLP, R2*: This document contains a description of the minimum lower layer protocol (MLLP), and its Release 2 extends the MLLP by providing support defining the minimal interpretation of reliable messaging.
- *HL7 Version 3 Standard: UML ITS Data Types, R1*: The UML data types specification binds the HL7 V3 data types to the UML/OCL kernel types to allow for formally correct OCL constrains on the V3 data types, and this assists in the implementation of the HL7 V3 data types.
- *HL7 Version 3 Standard: XML ITS Data Types, R1*: This specification defines the representation of HL7 V3 data types in XML.
- *HL7 Version 3 Standard: XML ITS Structures, R1*: This document defines the representation of HL7 V3 messages in XML, including the method to derive XML data type definitions (DTD)s.

HL7 V3 Messages: Administrative Domains (Representative Examples)

- *HL7 Version 3 Standard: Accounting and Billing, Release 2*: This document provides specifications for the creation and management of patient billing messages designed for the purpose of aggregating financial transactions that will be submitted as claims or invoices for reimbursement
- *HL7 Version 3 Standard: Claims and Reimbursement, R3*: This document provides specifications for generic claims, pharmacy

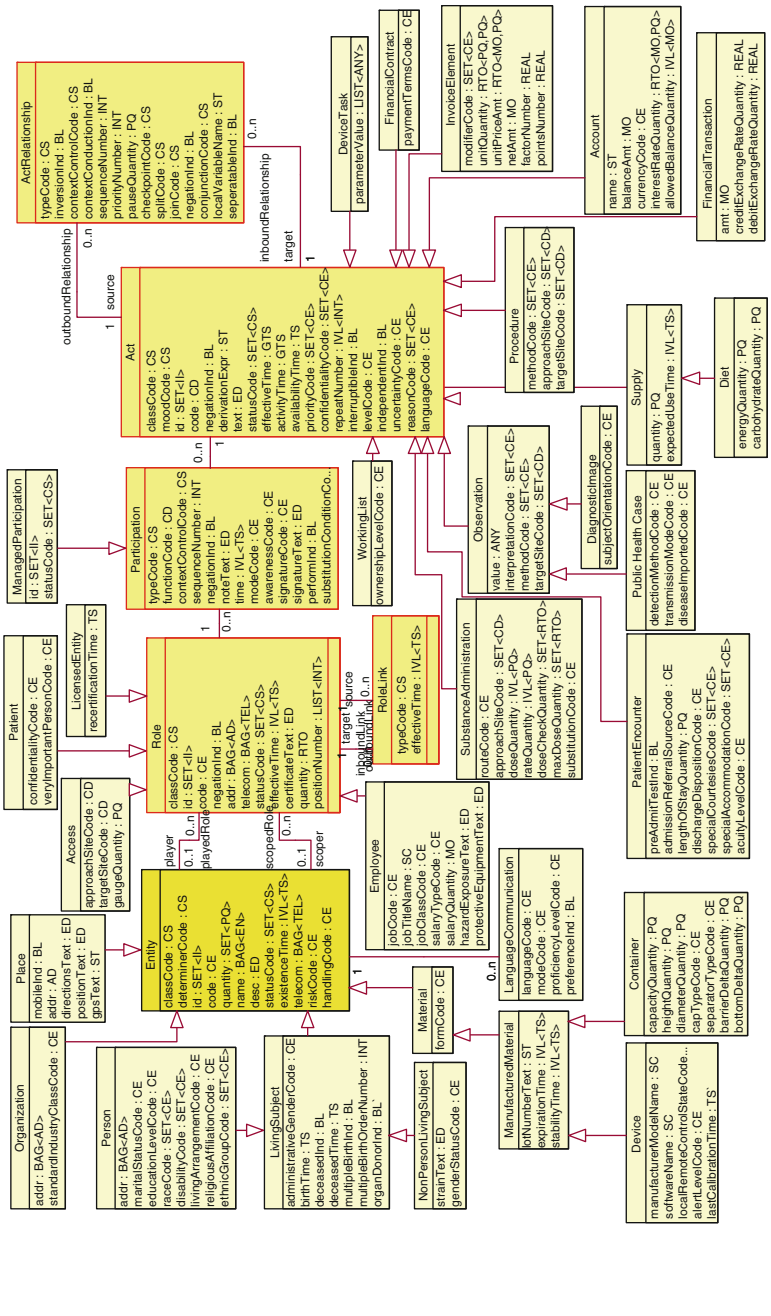


Fig. 7.2 HL7 Reference Information Model with the core classes highlighted

claims, preferred accommodation claims, physician reimbursement, oral health vision care, and hospital claims for eligibility, authorization, coverage extension, predetermination, invoice adjudication, and payment advice. The statement of financial activity (SOFA), in Release 3 of this document, added the claims messaging support for physician, oral health vision care, and hospital claims.

- *HL7 Version 3 Standard: Drug Stability Reporting, R1*: The Drug Stability Refined Message Information Model and Hierarchical Message Type capture pertinent to the drug stability testing process. Drug stability testing is required in the United States and in other countries as a component of the regulatory process. This testing verifies the correctness of a manufacturer's claims related to the stability (i.e., the ability to be stored over time without losing its therapeutic effectiveness) of a pharmaceutical product.
- *HL7 Version 3 Standard: Master File/Registry Infrastructure, R1*: This standard addresses the communications environments that are considered to be in common for all HL7 version 3 messaging implementations.
- *HL7 Version 3 Standard: Medical Records, R2 (Draft Standard for Trial Use (DSTU))*: The Medical Records DSTU addresses the information requirements for the management of clinical documents and any associated master files. Release 2 DSTU of this document adds queries to the current standard.
- *HL7 Version 3 Standard: Personnel Management, R1*: This document provides the modeling for provider and organization messages that are required to support registry-related messaging.
- *HL7 Version 3 Standard: Scheduling, R1*: This standard describes a messaging architecture for the notification of scheduling information from a scheduling system to auxiliary systems.

HL7 V3 Messages: Clinical Domains (Representative Examples)

- *HL7 Version 3 Standard: Care Provision; Professional Services, R1 (DSTU)*: The Care Provision domain addresses the information that is needed for the ongoing care of individuals, populations, and families.
- *HL7 Version 3 Standard: Care Structures Topic, R1 (DSTU)*: The Care Provision domain addresses the information that is needed for the ongoing care of individuals, populations, and other targets of care. The "Act of Care Provision" is the recording of a process that defines the responsibility for supplying support to the target of care. It is a statement of SUPERVISION, MANAGEMENT, and CUSTODY. The Care Structures Topic defines the many UML class diagrams, called care structures, which model the information pertinent to the ongoing provision of care.
- *HL7 Version 3 Standard: Clinical Genomics, Pedigree, R1*: The Pedigree Topic represents family history information. Several family history applications are in use by healthcare professionals and patients (including the Surgeon General's Family History tool), each having their own proprietary data format. This variation in knowledge representation makes it difficult to exchange information between and among these programs. By using this HL7 standard, users of disparate family history applications will be able to exchange an individual's family history information.
- *HL7 Version 3 Standard: Implantable Device Cardiac: Follow-up Device Summary, R1*: This standard models information related to the follow-up of patients who have received an implantable cardiac device (pacemaker, defibrillator, etc.). The information contains a subset of device observations, current device therapy settings, and device diagnostic information.
- *HL7 Version 3 Standard: Individual Case Safety Report, R1*: This document includes standards developed for the reporting of regulated information that extends outside the context of clinical trials. The current document contains messages which contain information related to adverse event notification and product stability reporting.
- *HL7 Version 3 Standard: Notifiable Condition Report, R1*: The Notifiable Condition Report captures the information needed to support "case reporting" between and among jurisdictional levels within the public health

system. This specification particularly focuses on the statutory/mandatory reporting of cases. It was specifically designed for the reporting of notifiable diseases or conditions as outlined by the Council of State and Territorial Epidemiologists and as adopted by the CDC. It also supports communications between field investigation team and their local health department.

- *HL7 Version 3 Standard: Regulated Product Submission, R1 (DSTU)*: The goal of the Regulated Product Submission message is to facilitate the FDA's processing and the review of submissions. The Regulated Product Submission Refined Message Information Model captures information required for the FDA's processing and review of regulatory submissions.
- *HL7 Version 3 Standard: Regulated Studies Annotated ECG, R1*: Clinical trials on candidate drug products often collect biological data from trial subjects as waveforms or algorithmic representations of those waveforms. After the data have been collected, derived measurements such as QT interval are then derived. This specification will be used to package annotated digital waveform data produced, for example, by an ECG analysis system for transmission from a trial sponsor or principal investigator to a regulatory agency. In no case would a waveform recording device be required to communicate its direct waveform readings using this specification.

Version 3 Rules/GELLO

GELLO is a standard expression language for decision support which is a specialization of the Object Constraint Language (OCL). OCL is developed by the Object Management Group (OMG) as a constraint for UML class models and as a query language for the information contained in these models. GELLO was designed to leverage the semantics of these HL7 models in combination with HL7 vocabulary and data types, for use in clinical decision support systems.

Arden Syntax

Arden is a "rules" specification that allows rules to be published independently of a computer system and subsequently imported into any computer systems.

CCOW/Visual Integration

Clinical Context Object Workgroup that allows users to experience an integrated computer-user session provides for the visual integration of applications. Messages are specified that flow between presentation-level applications that make more uniform the user identifier, patient identifier, and/or observation identifier across multiple applications for a "single-sign-on" and "single-patient-look-up" user experience.

Claims Attachments

Standard electronic attachments to a healthcare claim are a means of electronically exchanging additional information for adjudication of a healthcare claim, prior authorization, referrals, or for public health reporting.

Clinical Document Architecture (CDA® a V3-Based Standard)

The CDA Release 2.0 provides an exchange model for clinical documents such as discharge summaries and progress notes. CDA documents can be displayed using XML-aware web browsers (e.g., Internet Explorer, Firefox, Safire, or Google Chrome) or on mobile devices such as cell phones. The CDA is used as the format for the instantiation of the CCD model which is a care summary record based on the data elements of ASTM E31's Continuity of Care Record (CCR). This has been leveraged by the United States government as one of the standards selected to support health information exchange.

Electronic Health Record/Personal Health Record

The HL7 EHR System Functional Model provides a set of functions that are often present in an electronic health record (EHR) system. The function list is described from a user perspective with the intent to enable consistent discussion of system functionality. This EHR model, through the creation of functional profiles, is a standardized description of functions desired or currently available in a given setting's (e.g., intensive care, cardiology, general surgery) electronic systems.

Structured Product Labeling (a V3-Based Standard)

Often known as the standard to represent the content of FDA "product labels," "package inserts," or "prescribing information," this document, previously printed on paper, contains the authorized published information that accompanies any medication licensed for use in the United States. The SPL specification is a document markup standard that specifies the structure and semantics of the headings for these labels and, where feasible, is consistent with the HL7 Clinical Document Architecture (CDA®).

ASTM E31

The American Society of Testing and Materials International (ASTM) is an American National Standards Institute (ANSI)-accredited SDO which creates standards for a broad set of domains from electricity to concrete to healthcare. ASTM Committee E31 on Healthcare Informatics develops standards related to the architecture, terminologies, storage, security, confidentiality, functionality, and communication of information used within healthcare and healthcare decision making, including patient-specific information and knowledge. Established in 1970, E31 has a current membership of approximately 300 members, with 3 technical subcommittees that have over 30 approved standards and additional draft standards. Information on subcommittee structure and portfolio of approved standards and work items under development are available from the list of subcommittees, standards, and work items at the ASTM E31 web site. Approved standards are published annually in the Annual Book of ASTM Standards, Volume 14.01. An example of successful ASTM E31 standards is a set of security standards, many of which have also been accepted as standards by ISO, ASTM 2087 which is a standard for Quality Indicators for Terminologies in Health Informatics which has also been adopted by ISO as ISO TS17117, and the Continuity of Care Record [2].

The Continuity of Care Record (CCR), an ASTM E31 standard, is a care summary record and is a core data set that is intended to be sent to the next healthcare provider whenever a patient is referred, transferred, or otherwise uses different clinics, hospitals, or other providers. The CCR is intended to protect physicians and other healthcare professionals from having to act "blindly" without the needed access to relevant patient information. It provides the necessary information to support continuity of care, toward reducing medical errors, toward achieving higher efficiency, and toward the creation of a better quality of care.

Several US presidents have called for greater interoperability of electronic medical records and personal health records. ASTM E 2369-05, Standard Specification for the Continuity of Care Record (CCR) represents a major step forward in assisting vendors and healthcare organizations in their search for simple and powerful tools that will help meet the presidents' objectives.

Numerous sponsoring organizations have supported the efforts of ASTM Subcommittee E31.28 on Electronic Health Records throughout this process, including the Massachusetts Medical Society, the Healthcare Information and Management Systems Society, the American Academy of Family Physicians (AAFP), the American Academy of Pediatrics, the American Medical Association, the Patient Safety Institute, the American Health Care Association, the National Association for the Support of

Long-Term Care, the Mobile Healthcare Alliance, the Medical Group Management Association, and the American College of Osteopathic Family Physicians.

DICOM

DICOM is the national electrical manufacturer's association's (NEMA) standard for representing and sharing digital images and the information and metadata needed to understand the image, its associated results, and patient-related information. In the United States and in many countries abroad, DICOM is used to transmit digital images in support of radiological practice [3].

The introduction of digital medical image sources in the 1970s and the use of computers in processing these images after their acquisition led the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) to create a joint committee in order to develop a standard method for the transmission of medical images along with any relevant associated information. This committee, chartered in 1983, went on to publish in 1985 the ACR/NEMA standards publication No. 300-1985. Prior to this standard, most devices stored images in a proprietary format and transferred files of these formats over a network or on removable media in order to perform image communication. While the initial versions of the ACR-NEMA effort (i.e., version 2.0 – published in 1988) created standardized terminology, an information structure, and unsanctioned file encoding, most of the promise of a standard method of communicating digital image information was not realized until the release of version 3.0 of the DICOM standard in 1993. The release of version 3.0 saw a name change to Digital Imaging and Communications in Medicine (DICOM).

The DICOM standard now specified a network protocol utilizing TCP/IP, defined the operation of service classes beyond the simple transfer of data, and created a mechanism for uniquely identifying information objects as they are acted upon across the network. DICOM was also structured as a multipart document in order to facilitate extension of the standard. Additionally, DICOM defined information objects not only for images but also for patients, studies, reports, and other categories of data. With the enhancements made in DICOM (version 3.0), the standard permitted the transfer of medical images in a multivendor environment and also facilitated the development and expansion of picture archiving and communication systems (PACS) and interfacing with medical information systems.

CDISC

The Clinical Data Interchange Standards Consortium (CDISC) is an international, nonprofit organization that develops and supports global data standards for medical research. CDISC is working actively with the NIH National Cancer Institute's Enterprise Vocabulary Services (EVS) to develop and support controlled terminology in several areas, notably CDISC's Study Data Tabulation Model (SDTM). SDTM is an international standard for clinical research data and is approved by the FDA as a standard electronic submission format [4].

CDISC SDTM undergoes an extensive process of definition, development, and review before it is stamped as ready for release. Terminology that has completed this process is tagged as "Production" and now includes some 50 SDTM code lists with about 2,200 terms covering demographics, interventions, findings, events, trial design, units, frequency, and ECG terminology. This terminology is maintained and distributed as part of NCI Thesaurus and is available for direct download from the CDISC SDTM directory on an NCI file transfer protocol (FTP) site.

CDISC also leads the Clinical Data Acquisition Standards Harmonization (CDASH) project, which develops clinical research study content standards in collaboration with 16 partner organizations including NCI. NCI EVS maintains and distributes CDASH-controlled terminology as part of NCI Thesaurus. More information regarding

this project is available on CDISC's CDASH web page. CDASH terminology is a subset of the SDTM terminology and is available for direct download from the CDISC CDASH directory on an NCI file transfer protocol (FTP) site.

CDISC also leads the Analysis Data Model (ADaM) project, which is meant to support efficient generation, replication, review, and submission of analysis results from clinical trial data. The NCI EVS maintains and distributes ADaM-controlled terminology as part of the NCI Thesaurus. The ADaM terminology is available for direct download from the CDISC or from the ADaM directory on an NCI file transfer protocol (FTP) site.

CDISC also leads the Standard for the Exchange of Nonclinical Data (SEND) project, which guides the organization, structure, and format of standard nonclinical tabulation data sets for interchange between organizations such as sponsors and CROs and for submission to a regulatory authority such as the FDA. NCI EVS maintains and distributes the SEND-controlled terminology as part of NCI Thesaurus. The SEND terminology is available for direct download from the CDISC SEND directory on an NCI file transfer protocol (FTP) site.

The CDISC New Term Request web page handles suggestions for both new terminology and changes to existing terminology. The CDISC Term Request Tracking Excel spreadsheet helps members of the CDISC community review and comment on all submitted requests.

OMG

The Object Management Group (OMG) is a standards development organization that has, as a component, a healthcare-specific mission for the development of a service-oriented architecture (SOA). This activity is a joint project between the OMG and the HL7 SDOs. These services are intended to be the backbone of communications in healthcare toward true interoperable data. The process of service creation and balloting of the services is still an ongoing project [5].

The Common Terminology Services (CTS) II specification was developed as an alternative to a common data structure. Instead of specifying what an external terminology must look like, HL7 has chosen to identify the common functional characteristics that an external terminology must be able to provide. As an example, an HL7-compliant terminology service will need to be able to determine whether a given concept code is valid within the particular resource. Instead of describing a table keyed by the resource identifier and concept code, the CTS specification describes an application programming interface (API) call that takes a resource identifier and concept code as input and returns a true/false value. Each terminology developer is free to implement this API call in whatever way is most appropriate for them.

The CTS specification is not designed to perform the following services:

- The current version of CTS is not intended to be a complete terminology service. The scope of CTS is restricted to the functionality needed to design, implement, and deploy an HL7 version 3–compliant software package.
- CTS is not intended to be a general-purpose query language. It is intended to specify only the specific services needed in the HL7 implementation.
- CTS II does not specify how the service is to be implemented. It is intentionally silent when it comes to service advertising and discovery, establishing and maintaining connections, and the delivery and routing of messages. It is assumed that a CTS implementation will use the underlying architecture that is most appropriate for the given implementation circumstances.

OMG is an international, open-membership, not-for-profit computer industry consortium. OMG task forces develop enterprise integration standards for a wide range of technologies, including real-time, embedded and specialized systems, analysis and design, architecture-driven modernization, and middleware and an even wider range of industries, including business modeling and integration, C4I, finance, government, healthcare, insurance, legal compliance, life sciences research, manufacturing technology, robotics, software-based communications, and space.

OMG's modeling standards, including the Unified Modeling Language™ (UML®) and Model Driven Architecture® (MDA®), enable powerful visual design, execution, and maintenance of software and other processes, including IT systems modeling and business process management. OCL is also a standard of OMG and is used to constrain UML models. OMG's middleware standards and profiles are based on the Common Object Request Broker Architecture (CORBA®) and support a wide variety of industries. More information about OMG can be found at www.omg.org.

OASIS

Organization for the Advancement of Structured Information Standards (OASIS) is a not-for-profit consortium that drives the development, convergence, and adoption of open standards for the global information society [6].

OASIS promotes industry consensus and produces worldwide standards for security, cloud computing, SOA, web services, the smart grid, electronic publishing, emergency management, and healthcare. OASIS open standards offer the potential to lower cost, stimulate innovation, grow global markets, and protect the right of free choice of technology.

OASIS members broadly represent the marketplace of public and private sector technology leaders, users, and influencers. The consortium has more than 5,000 participants representing over 600 organizations and individual members in 100 countries.

OASIS is distinguished by its transparent governance and operating procedures. Members themselves set the OASIS technical agenda, using a lightweight process expressly designed to promote industry consensus and unite disparate efforts. Completed work is ratified by open ballot. Governance is accountable and unrestricted. Officers of both the OASIS Board of Directors and Technical Advisory Board are chosen by democratic election to serve 2-year terms. Consortium leadership is based on individual merit and is not tied to financial contribution, corporate standing, or special appointment.

IHTSDO

The International Health Terminology Standards Development Organization (IHTSDO) is an international not-for-profit organization based in Denmark. IHTSDO acquires (from contributors), owns, and administers the rights to SNOMED CT and other health terminologies and related standards [7].

The purpose of IHTSDO is to develop, maintain, promote, and enable the uptake and correct use of its terminology products in health systems, services, and products around the world and undertake any or all activities incidental and conducive to achieving the purpose of the association for the benefits of the members.

The IHTSDO seeks to improve the health of humankind by fostering the development and use of suitable standardized clinical terminologies, notably SNOMED CT, in order to support safe, accurate, and effective exchange of clinical and related health information. The focus is on enabling the implementation of semantically accurate health records that are interoperable. Support to association members and licensees is provided on a global basis, allowing the pooling of resources to achieve shared benefits.

The Objects of the association are to:

(a) Enhance the health of humankind by facilitating better health information management
(b) Contribute to improved delivery of care by clinical and social care professions
(c) Facilitate the accurate sharing of clinical and related health information and the semantic interoperability of health records
(d) Encourage global collaboration and cooperation with respect to the ongoing improvement of the terminology products
(e) Provide the foregoing on a globally coordinated basis, thereby enabling the members and the related organizations within their territories to pool resources and share benefits relating to the development and maintenance of, and their utilization of, and reliance upon the terminology products.

For more detailed information regarding SNOMED CT, please see the chapter on "Terminology Implementation."

CEN TC 251

The European Committee for Standardization (CEN) is the pan-Europe committee that creates and ballots standards. Technical Committee 251 concerns itself with health informatics standardization [8]. Many important terminological standards have been published by CEN including:

Standard reference	Title
CEN/TR 15212:2006	Health informatics – Vocabulary – Maintenance procedure for a web-based terms and concepts database
CEN/TR 15253:2005	Health informatics – Quality of service requirements for health information interchange
CEN/TR 15299:2006	Health informatics – Safety procedures for identification of patients and related objects
CEN/TR 15300:2006	Health informatics – Framework for formal modeling of healthcare security policies
CEN/TR 15640:2007	Health informatics – Measures for ensuring the patient safety of health software
CEN/TS 14822-4:2005	Health informatics – General purpose information components – Part 4: Message headers
CEN/TS 15127-1:2005	Health informatics – Testing of physiological measurement software – Part 1: General
CEN/TS 15260:2006	Health informatics – Classification of safety risks from health informatics products
CEN/TS 15699:2009	Health informatics – Clinical knowledge resources – Metadata
CR 12161:1995	A method for defining profiles for healthcare
CR 12587:1996	Medical Informatics – Methodology for the development of healthcare messages
CR 1350:1993	Investigation of syntaxes for existing interchange formats to be used in healthcare
CR 13694:1999	Health Informatics – Safety and Security Related Software Quality Standards for Healthcare (SSQS)
CR 14301:2002	Health informatics – Framework for security protection of healthcare communication
CR 14302:2002	Health informatics – Framework for security requirements for intermittently connected devices
EN 1064:2005 + A1:2007	Health informatics – Standard communication protocol – Computer-assisted electrocardiography
EN 1068:2005	Health informatics – Registration of coding systems
EN 12251:2004	Health informatics – Secure User Identification for Health Care – Management and Security of Authentication by Passwords
EN 12264:2005	Health informatics – Categorial structures for systems of concepts
EN 12381:2005	Health informatics – Time standards for healthcare specific problems
EN 12435:2006	Health informatics – Expression of results of measurements in health sciences
EN 13606-1:2007	Health informatics – Electronic health record communication – Part 1: Reference model
EN 13606-2:2007	Health informatics – Electronic health record communication – Part 2: Archetypes interchange specification
EN 13606-3:2008	Health informatics – Electronic health record communication – Part 3: Reference archetypes and term lists
EN 13606-4:2007	Health informatics – Electronic health record communication – Part 4: Security
EN 13609-1:2005	Health informatics – Messages for maintenance of supporting information in healthcare systems – Part 1: Updating of coding schemes
EN 13940-1:2007	Health informatics – System of concepts to support continuity of care – Part 1: Basic concepts
EN 14463:2007	Health informatics – A syntax to represent the content of medical classification systems – ClaML
EN 14484:2003	Health informatics – International transfer of personal health data covered by the EU data protection directive – High level security policy
EN 14485:2003	Health informatics – Guidance for handling personal health data in international applications in the context of the EU data protection directive
EN 14822-1:2005	Health informatics – General purpose information components – Part 1: Overview

Standard reference	Title
EN 14822-2:2005	Health informatics – General purpose information components – Part 2: Non-clinical
EN 14822-3:2005	Health informatics – General purpose information components – Part 3: Clinical
EN 15521:2007	Health informatics – Categorial structure for terminologies of human anatomy
EN 1614:2006	Health informatics – Representation of dedicated kinds of property in laboratory medicine
EN 1828:2002	Health informatics – Categorial structure for classifications and coding systems of surgical procedures
EN ISO 10781:2009	Electronic Health Record- System Functional Model, Release 1.1 (ISO 10781:2009)
EN ISO 11073-10101:2005	Health informatics – Point-of-care medical device communication – Part 10101: Nomenclature (ISO/IEEE 11073-10101:2004)
EN ISO 11073-10201:2005	Health informatics – Point-of-care medical device communication – Part 10201: Domain information model (ISO/IEEE 11073-10201:2004)
EN ISO 11073-10404:2011	Health informatics – Personal health device communication – Part 10404: Device specialization – Pulse oximeter (ISO/IEEE 11073-10404:2010)
EN ISO 11073-10407:2011	Health informatics – Personal health device communication – Part 10407: Device specialization – Blood pressure monitor (ISO/IEEE 11073-10407:2010)
EN ISO 11073-10408:2011	Health informatics – Personal health device communication – Part 10408: Device specialization – Thermometer (ISO/IEEE 11073-10408:2010)
EN ISO 11073-10415:2011	Health informatics – Personal health device communication – Part 10415: Device specialization – Weighing scale (ISO/IEEE 11073-10415:2010)
EN ISO 11073-10417:2011	Health informatics – Personal health device communication – Part 10417: Device specialization – Glucose meter (ISO/IEEE 11073-10417:2010)
EN ISO 11073-10471:2011	Health Informatics – Personal health device communication – Part 10471: Device specialization – Independent living activity hub (ISO/IEEE 11073-10471:2010)
EN ISO 11073-20101:2005	Health informatics – Point-of-care medical device communication – Part 20101: Application profiles – Base standard (ISO/IEEE 11073-20101:2004)
EN ISO 11073-20601:2011	Health informatics – Personal health device communication – Part 20601: Application profile – Optimized exchange protocol (ISO/IEEE 11073-20601:2010)
EN ISO 11073-30200:2005	Health informatics – Point-of-care medical device communication – Part 30200: Transport profile – Cable connected (ISO/IEEE 11073-30200:2004)
EN ISO 11073-30300:2005	Health informatics – Point-of-care medical device communication – Part 30300: Transport profile – Infrared wireless (ISO/IEEE 11073-30300:2004)
EN ISO 12052:2011	Health informatics – Digital imaging and communication in medicine (DICOM) including workflow and data management (ISO 12052:2006)
EN ISO 12967-1:2011	Health informatics – Service architecture – Part 1: Enterprise viewpoint (ISO 12967-1:2009)
EN ISO 12967-2:2011	Health informatics – Service architecture – Part 2: Information viewpoint (ISO 12967-2:2009)
EN ISO 12967-3:2011	Health informatics – Service architecture – Part 3: Computational viewpoint (ISO 12967-3:2009)
EN ISO 13606-5:2010	Health informatics – Electronic health record communication – Part 5: Interface specification (ISO 13606-5:2010)
EN ISO 18104:2003	Health Informatics – Integration of a reference terminology model for nursing (ISO 18104:2003)
EN ISO 18812:2003	Health informatics – Clinical analyser interfaces to laboratory information systems – Use profiles (ISO 18812:2003)
EN ISO 21090:2011	Health Informatics – Harmonized data types for information interchange (ISO 21090:2011)
EN ISO 21549-1:2004	Health informatics – Patient healthcard data – Part 1: General structure (ISO 21549-1:2004)
EN ISO 21549-2:2004	Health informatics – Patient healthcard data – Part 2: Common objects (ISO 21549–2:2004)
EN ISO 21549-3:2004	Health informatics – Patient healthcard data – Part 3: Limited clinical data (ISO 21549-3:2004)

Standard reference	Title
EN ISO 21549-4:2006	Health informatics – Patient healthcard data – Part 4: Extended clinical data (ISO 21549-4:2006)
EN ISO 21549-5:2008	Health informatics – Patient healthcard data – Part 5: Identification data (ISO 21549-5:2008)
EN ISO 21549-6:2008	Health informatics – Patient healthcard data – Part 6: Administrative data (ISO 21549-6:2008)
EN ISO 21549-7:2007	Health informatics – Patient healthcard data – Part 7: Medication data (ISO 21549-7:2007)
EN ISO 21549-8:2010	Health informatics – Patient healthcard data – Part 8: Links (ISO 21549-8:2010)
EN ISO 27799:2008	Health informatics – Information security management in health using ISO/IEC 27002 (ISO 27799:2008)
ENV 12443:1999	Medical Informatics – Healthcare Information Framework (HIF)
ENV 12537-1:1997	Medical informatics – Registration of information objects used for EDI in healthcare – Part 1: The Register
ENV 12610:1997	Medical informatics – Medicinal product identification
ENV 12611:1997	Medical informatics – Categorial structure of systems of concepts – Medical devices
ENV 12612:1997	Medical informatics – Messages for the exchange of healthcare administrative information
ENV 13607:2000	Health informatics – Messages for the exchange of information on medicine prescriptions
ENV 13608-1:2000	Health informatics – Security for healthcare communication – Part 1: Concepts and terminology
ENV 13608-2:2000	Health informatics – Security for healthcare communication – Part 2: Secure data objects
ENV 13608-3:2000	Health informatics – Security for healthcare communication – Part 3: Secure data channels
ENV 13609-2:2000	Health informatics – Messages for maintenance of supporting information in healthcare systems – Part 2: Updating of medical laboratory-specific information
ENV 13730-1:2001	Health informatics – Blood transfusion related messages – Part 1: Subject of care related messages
ENV 13730-2:2002	Healthcare Informatics – Blood transfusion related messages – Part 2: Production related messages (BTR-PROD)

Table 7.1 European Standards by their identifier and their descriptive name

The *European Committee for Standardization* (CEN) is a business facilitator in Europe, removing trade barriers for European industry and consumers. Its mission is to foster the European economy in global trading and the welfare of European citizens and the environment. Through its services, it provides a platform for the development of European Standards and other technical specifications.

CEN is a major provider of European Standards and technical specifications. It is the only recognized European organization according to Directive 98/34/EC for the planning, drafting, and adoption of European Standards in all areas of economic activity with the exception of electrotechnology (CENELEC) and telecommunication (ETSI).

CEN's 31 national members work together to develop voluntary European Standards (ENs).

These standards have a unique status since they also are national standards in each of its 31 member countries. With one common standard in all these countries and every conflicting national standard withdrawn, a product can reach a far wider market with much lower development and testing costs. ENs help build a European internal market for goods and services and position Europe in the global economy. More than 60,000 technical experts as well as business federations, consumer, and other societal interest organizations are involved in the CEN network that reaches over 480 million people.

In a globalized world, the need for international standards simply makes sense. CEN TC 251 standards can be fast tracked into ISO as specified in the Vienna Agreement, which was ratified in 1991. In this way, the European community has been a positive world benefactor of health informatics standards (Table 7.1).

ISO TC 215

Standardization in the field of information for health and health information and communications technology (ICT) to promote interoperability between independent systems, to enable compatibility and consistency for health information and data, as well as to reduce duplication of effort and redundancies [9].

The domain of ICT for health includes but is not limited to:

- Healthcare delivery
- Disease prevention and wellness promotion
- Public health and surveillance
- Clinical research–related to health service

Total number of published ISO standards related to the TC and its SCs (number includes updates):	89
Number of published ISO standards under the direct responsibility of TC 215 (number includes updates):	89
Participating countries:	32
Observing countries:	20

The organizational structure of ISO TC 215 is comprised of the following working groups and committee:

The working group 3 is the terminology-related working group; however, there are terminological constraints and requirements that are needed for the business of each of the other working groups, and therefore strong liaisons remain between working group 3 and the other working groups. The Healthcare Information and Management Systems Society (HIMSS) serves as the current secretariat of ISO TC 215 (Table 7.2).

ISO/TC 215 was established during 1998 with the following scope:

> Standardization in the field of information for health, and Health Information and Communications Technology (ICT) to achieve compatibility and interoperability between independent systems. Also, to ensure compatibility of data for comparative statistical purposes (e.g., classifications), and to reduce duplication of effort and redundancies

At its first meeting on August 25/26, 1998, ISO/TC 215 established four working groups (1–4). The purpose of working group 3 is to develop standards in the area of semantic content.

Table 7.2 ISO Technical Committee 215 working groups by their identifiers and names

TC 215/CAG 1	Executive council, harmonization, and operations
TC 215/WG 1	Data structure
TC 215/WG 2	Data interchange
TC 215/WG 3	Semantic content
TC 215/WG 4	Security
TC 215/WG 6	Pharmacy and medicines business
TC 215/WG 7	Devices
TC 215/WG 8	Business requirements for electronic health records
TC 215/WG 9	SDO harmonization

The scope of ISO/TC 215 working group 3 is to develop standards to support the representation of (1) health concepts. These standards include formal models of representation and description of health concepts, principles of their organization within (2) terminologies and their related systems (including controlled clinical terminologies and classifications), and issues concerning the context of their use in electronic health records.

WG3 will develop or adopt standards, including metavocabularies, to address:

- Structure (including semantic models)
- Development
- Function
- Implementation
- Use (including compatibility with other relevant information models, such as health records, messaging)
- Distribution
- Evaluation
- Maintenance (including the editing environment, updating, managing change) of terminologies and their related systems. It does not include the creation, endorsement, or maintenance of detailed terminology contents.

This scope also includes mechanisms of aggregation, including mapping between a terminology and a statistical classification.

1. Health concepts in this context are considered to include all disciplines and concepts necessary to maintain health (social, environmental, physiological, and mental) and prevent or treat ill health.

2. Terminologies and related systems in this context are considered to include all paper or electronic-based systems of concepts designed to record and/or categorize information, along with the phrases used to express them in different natural languages.

The scope of the ISO TC WG3 work program is governed by these principles, that WG3 will:

- Perform all its activities within the scope of work determined by the governing Technical Committee ISO/TC 215
- Create a framework of standards that enables health information to be created, used, and shared across any and all boundaries including systems, jurisdictions, disciplines, languages, and professions
- Address all health information standards work that fall within the scope of ISO/TC 215 WG 3
- Maintain liaison with standards bodies and other organizations
- Establish and maintain liaison and, where necessary, develop integrated standards, with other ISO/TC 215 working groups

The standards developed will:

- Employ existing modeling notations
- Take into account current regional and national work in all activities of WG3
- Not be limited to application within computerized systems
- Be clear, relevant, needed, and implementable

An example of an important terminological standard generated by this working group is ISO/TS 17117 – Health informatics – Criteria for the categorization and evaluation of terminological systems, originally authored by Dr. Peter L. Elkin. This standard defines the desirable qualities of a terminology and how to best evaluate a terminology within its scope and purpose. TS 17117 is a direct descendant of ASTM's standard on Quality Indicators for Healthcare Terminologies, ASTM 2087. For more information regarding the quality criteria for healthcare terminologies, please see the chapter on the "Theoretical Foundations of Terminology."

NCHS

The mission of the National Center for Health Statistics (NCHS) is to provide statistical information that will guide actions and policies to improve the health of the American people. As the nation's principal health statistics agency, NCHS leads the way with accurate, relevant, and timely data [10].

The National Center for Health Statistics is a rich source of information about America's health. As the nation's principal health statistics agency, NCHS compiles statistical information to guide actions and policies to improve the health of our people. NCHS provides a unique public resource for health information – a critical element of public health and health policy.

Our health statistics allow us to:

- Document the health status of the population and of important subgroups
- Identify disparities in health status and use of healthcare by race or ethnicity, socioeconomic status, region, and other population characteristics
- Describe our experiences with the healthcare system
- Monitor trends in health status and healthcare delivery
- Identify health problems
- Support biomedical and health services research
- Provide information for making changes in public policies and programs
- Evaluate the impact of health policies and programs

Working with partners throughout the health community, NCHS uses a variety of approaches to efficiently obtain information from the sources most able to provide information. NCHS collects data from birth and death records, medical records, interview surveys, and through direct physical exams and laboratory testing. NCHS is a key element of our national public health infrastructure, providing important surveillance information that helps identify and address critical health problems.

At NCHS, information is at the core of our mission. The National Committee on Vital and

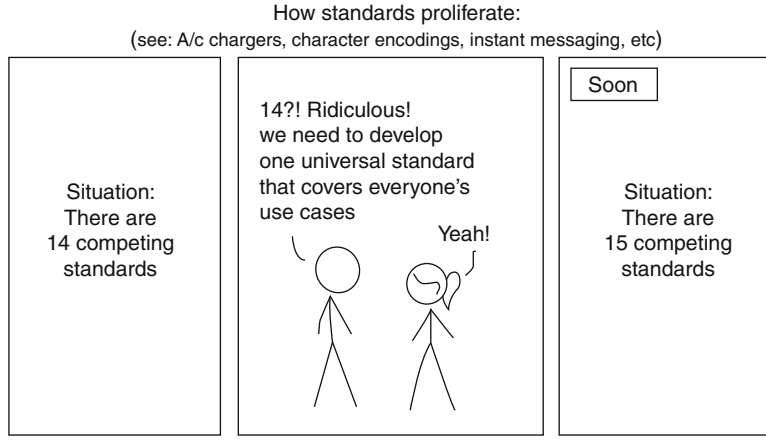


Fig. 7.3 Proliferation of health informatics standards

Health Statistics was established by congress to serve as an advisory body to the Department of Health and Human Services on health data, statistics, and national health information policy. It fulfills important review and advisory functions relative to health data and statistical problems of national and international interest, stimulates or conducts studies of such problems, and makes proposals for improvement of the nation's health statistics and information systems.

The Public Health Data Standards Consortium is an important vehicle for promoting standardization of information on health and healthcare. The National Center for Health Statistics (NCHS) was instrumental in establishing the Public Health Data Standards Consortium (Consortium) in 1999. The Consortium, which incorporated as a not-for-profit organization in 2003, is a national nonprofit member-based partnership of federal, state, and local health agencies, national and local professional associations, and public and private sector organizations and individuals.

Conclusions

There is an often used parable that the problem with standards is that there are so many to choose from (see Fig. 7.3). The standards listed here all evolved for a reason. It seems that one unifying health informatics standard is not likely to arise and indeed may be inappropriate. Each of these standards was created with a scope and purpose in mind. Within that scope, the standard may perform better than any unifying model would perform. Therefore, guidance of which standards should be used within which subdomains of health and healthcare is more likely to become the accepted method for applying standards. We commend the health informatics standards community for the multidecade effort that has led to the level of terminological standardization at our command today. This is the result of a global effort that reflects the intellect of many individuals from across the globe. I am inspired by your brilliance and awed by the dogged effort which you applied toward the betterment of human health. That notwithstanding, there is much work left to be accomplished. Each reader is encouraged to lend their talents to the global community, creating and disseminating health informatics standards.

Questions

1. What SDO publishes TS17117?
   (a) ASTM
   (b) HL7
   (c) DICOM
   (d) ISO
2. What SDO publishes EN 12611?
   (a) IHTSDO
   (b) CEN TC 251
   (c) ISO TC 215
   (d) OASIS

3. The national electrical manufacturer's association is associated with which standard?
   (a) DICOM
   (b) Reference Information Model
   (c) SNOMED CT
   (d) Quality Indicators for HIT
4. A standard for developing models in the healthcare domain includes?
   (a) HL7
   (b) SNOMED CT
   (c) OMG
   (d) All of the above
5. All of the following organizations create content for healthcare except?
   (a) SNOMED CT
   (b) HL7
   (c) ISO TC 215
   (d) ASTM
6. NCHS is a part of which US Government Agency?
   (a) FDA
   (b) CDC
   (c) NIH
   (d) ONC
7. Which pair(s) of standards is/are related?
   (a) ASTM 2087 and ISO TS 17117
   (b) HL7 RIM and CEN 13606
   (c) ASTM CCR and HL7 CCD
   (d) a and c
   (e) a, b, and c
8. Which standard concerns itself principally with the human computer interaction?
   (a) HL7
   (b) ASTM
   (c) CDISC
   (d) OASIS
9. Which standard represents methods for the transport layer standards?
   (a) OASIS
   (b) HL7
   (c) CDISC
   (d) ASTM
10. Which SDO provides all of the knowledge needed for health and healthcare?
    (a) HL7
    (b) ASTM
    (c) ISO TC 215
    (d) CEN TC 251
    (e) None of the above

References

1. www.hl7.org.
2. http://www.astm.org/COMMIT/COMMITTEE/E31.htm.
3. http://medical.nema.org/.
4. www.cdisc.org.
5. www.omg.org.
6. http://www.oasis-open.org/standards.
7. www.ihtsdo.org.
8. http://www.cen.eu/CEN/Sectors/TechnicalCommitteesWorkshops/CENTechnicalCommittees/Pages/Standards.aspx?param=6232&title=CEN/TC±251.
9. http://www.iso.org/iso/iso_technical_committee?commid=54960.
10. http://www.cdc.gov/nchs/index.htm.

Implementations of Terminology

8

Peter L. Elkin, Mark Samuel Tuttle, Marjorie Rallins, Jennifer Trajkovski, Elizabeth Lumakovska, and Steven H. Brown

This chapter will not present a complete accounting of all terminology implementations. However, we will provide a survey of some of the more highly adopted terminologies. We will describe their scope and purpose, and we will discuss their implementation and the risks and benefits associated with employing each of these terminologies. In so doing we will provide examples of terminologies that can be used by healthcare informaticians to represent health knowledge and to use those representations for clinical decision support.

In this chapter we will describe:

- Unified Medical Language System (UMLS)
- NCI EVS
- WHO Family of Classifications (ICD and specifically ICD9-CM and ICF and ICHI)
- CPT
- LOINC
- SNOMED CT
- NDF-RT
- RxNorm
- ICNP
- OBO and GO

During our discussions of these terminologies, we will mention related terminologies. In this way, we hope to give the readers a more complete and balanced view of the field without exhaustively describing all international terminological implementations.

The background gained in Chaps. 3 and 4 will serve the student well in understanding the information presented in this chapter.

P.L. Elkin, M.D., MACP, FACMI (✉)
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

M.S. Tuttle, AB, BE, FACMI
Apelon, Ridgefield, CT, USA

M. Rallins, DPM, MA • J. Trajkovski, MJ, RHIT, CHC
Performance Improvement, American Medical Association, 515 N. State St, Chicago, IL, USA

E. Lumakovska, MPA
CPT Medical Informatics and Healthcare Strategy, American Medical Association, 515 N. State St, Chicago, IL, USA

S.H. Brown, M.D., MS, FACMI
Department of Veterans Affairs and Vanderbilt, Department of Biomedical Informatics, Nashville, TN, USA

Unified Medical Language System (UMLS)

The purpose of the National Library of Medicine®'s Unified Medical Language System (UMLS) is to provide knowledge sources and tools that serve to facilitate the development of computer applications that act as if they "understand" the meaning of the language of biology, medicine, and health [1]. The UMLS provides information for systems' developers as well as applications that support searching and reporting functions aimed at the less technical user. The major driving forces, from the National Library of Medicine of the US National Institutes of Health behind the UMLS's development were NLM Director Donald

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_8, © Springer-Verlag London 2012

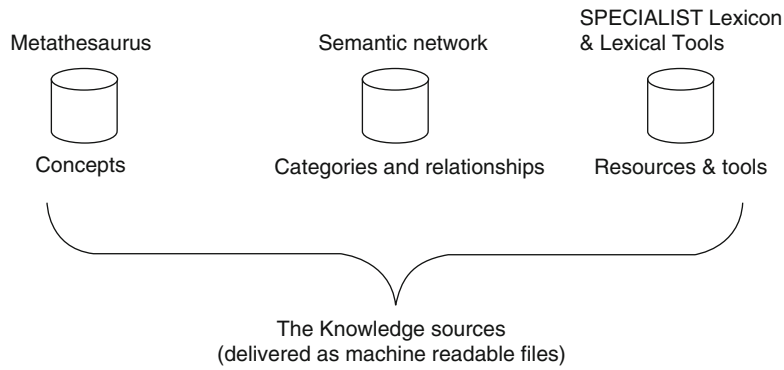


Fig. 8.1 UMLS knowledge sources

Lindberg, M.D., and Betsy Humphreys, MLS Over time this development represented a convergence of the talents of clinicians, informaticians, librarians, linguists, and computer scientists specializing in computational linguistics.

There are three UMLS knowledge sources (see Fig. 8.1):

- The Metathesaurus®, which is a compendium of terminologies and serves to provide over one million biomedical concepts from over 100 source vocabularies
- The Semantic Network, which defines 133 broad categories and 54 relationships between categories for representing knowledge in the biomedical domain
- The SPECIALIST Lexicon and Lexical Tools, which contains lexical information and a set of computer programs that perform natural language processing

The UMLS install and customization program is distributed with a set of lexical tools and MetamorphoSys.

The UMLS Metathesaurus

The Metathesaurus is a large, multilingual vocabulary database that contains information about biomedical and health-related concepts, their various synonyms, and the relationships between concepts. The Metathesaurus is built from the electronic versions of many different thesauri, terminologies, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing of terminologies and cataloging the biomedical literature, and/or for basic, clinical, and health services research.

Table 8.1 Examples of representation of the same concept by a set of terminologies

Term	Terminology
Atrial fibrillation	ICD9-CM
AF	NCI Thesaurus
AFib	MedDRA
Atrial fibrillation (disorder)	SNOMED Clinical Terms
Atrium; fibrillation	ICPC2-ICD10 Thesaurus

The Metathesaurus currently contains over five million terms, or names, organized by meaning into concepts and assigned with a unique identifier. The data in the Metathesaurus is stored in a set of relational tables and files. A free tool is distributed with the UMLS Metathesaurus that can be installed locally using MetamorphoSys.

The UMLS Metathesaurus is not a vocabulary. It contains many vocabularies, many of which are in themselves national or international standards, and the Metathesaurus helps to create mappings between and among these vocabularies, but it has never been the intent to replace them or their function.

Over 100 vocabularies, code sets, and thesauri, or "source vocabularies" are brought together to produce the Metathesaurus. Terms from each source vocabulary are organized by meaning and assigned a concept unique identifier (CUI), which they called the "name that never changes."

Sixty-two percent of the Metathesaurus's source vocabularies are provided in English. The Metathesaurus also contains terms from 17 other languages such as Spanish, French, Dutch, Italian, Japanese, and Portuguese.

Table 8.1 shows some of the terms that are associated with the CUI is the identifier associated with the concept. The source vocabulary that

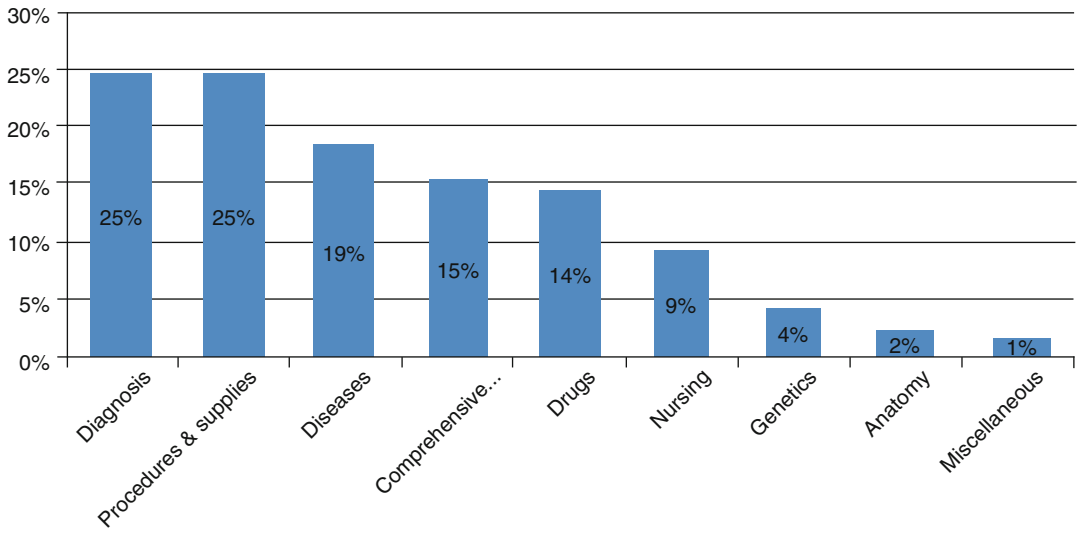


Fig. 8.2 The percentage of the Metathesaurus that is classified into a set of vocabulary categories

contributed each term is listed next to it. Often a source vocabulary will contribute more than one term to a concept (identified by a concept unique identifier (CUI)).

The UMLS Metathesaurus provides all of the original data from the source vocabulary including unique identifiers, definitions, or term spelling variants and organizes the data into a common format.

Figure 8.2 shows the percentage of the Metathesaurus that is classified into a set of vocabulary categories. This shows the broad distribution of category types within the UMLS.

At the time of installation, one can use the program MetamorphoSys to create vocabulary subsets. These subsets can be by language, semantic type or by vocabulary, for example.

Unique Identifiers in the Metathesaurus

When a concept is added to the Metathesaurus, it receives a concept unique identifier. The Metathesaurus structure has four levels of specification:

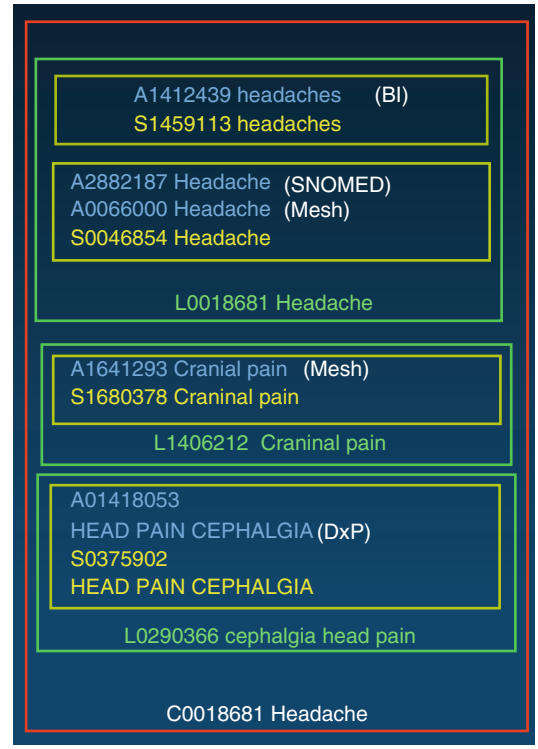


Fig. 8.3 Terms and strings associated with the concept Headache and its CUI

Concept Unique Identifiers (CUI)

A concept connotes one meaning, and the CUI is the representation of that meaning. A meaning can have many different names associated with the concept. A key goal in the construction of the Metathesaurus is to understand the intended meaning of each name in each source vocabulary and to link all the names from all of the source vocabularies that mean the same thing (i.e., are

synonyms). CUIs begin with the letter C and are followed by seven numbers. In the example shown in Fig. 8.3, the CUI is C0018681 and represents the concept of *Headache*.

Lexical (Term) Unique Identifiers (LUI)

LUIs link strings together that are lexical variants. Lexical variants are detected using the lexical variant generator (LVG) program, one of the tools in the UMLS lexical tools. LUIs begin with the letter L and are followed by seven numbers. In the example, taken from the UMLS website, on the right, there are three lexical variants; each has a unique LUI.

String Unique Identifiers (SUI)

Each unique string (set of characters) in each language (English, French, Spanish, etc.) in the Metathesaurus has a unique and permanent string unique identifier (SUI). Any variation in character set, upper–lower case, or punctuation difference is considered a separate (unique) string and is assigned a separate SUI. SUIs begin with the letter S followed by seven numbers. In the example on the right there are four unique strings, each assigned a different SUI.

Atom Unique Identifiers (AUI)

The basic building blocks or "atoms" with which the Metathesaurus is constructed are the concept names or strings contained within each of the source vocabularies. Every occurrence of a string in each source vocabulary is assigned an atom unique identifier (AUI). If exactly the same string is identified to have multiple occurrences from within the same vocabulary, for example, as an alternate name for different concepts (ambiguous abbreviations such as "MS"), an AUI is assigned for each occurrence. AUIs begin with the letter A followed by seven numbers. In the example above, there are five unique strings derived from five vocabulary sources and therefore are assigned five different AUIs. The abbreviation for the source vocabulary that contributed each string is denoted within the parentheses shown immediately after the printed string.

The Metathesaurus consists of 40 files including data, metadata, and index file types. The data files listed below contain information obtained from the source vocabularies in the rich release format. Concept unique identifiers (CUI) link data related to a concept across the individual files. Table 8.2 illustrates the information that populates each data file type.

Table 8.2 Files from the UMLS and what information they contain

Metadata file name	Contents
MRCONSO.RRF	Names, synonyms, terms, term types, codes
MRREL.RRF	Relationships
MRHIER.RRF	Hierarchies
MRSAT.RRF	Attributes
MRDEF.RRF	Definitions
MRMAP.RRF	Mappings
MRSMAP.RRF	Simplified mappings
MRSTY.RRF	Semantic types

In addition to data files, two other file types are released with each Metathesaurus version.

Index files are produced to help developers build applications that search for specific words or groups of words from within the content. For example, the index file MRXNW_ENG.RRF connects words to all related strings and links CUIs to concept identifiers.

Metadata files contain information about each specific release of the Metathesaurus including its sources and the files contained within the release. For example, MRFILES.RRF contains a listing of all the files provided in any Metathesaurus subset with a brief description and a listing of all rows and columns contained in that release.

Subsets may include files with a size of 0 or more bytes. Subsets can exclude some files. For example, MRXW_DUT.RRF will only be included in a subset that contains terms from the Dutch language which may be more interesting to you if you live in the Netherlands.

The Semantic Network

The Semantic Network consists of both a set of semantic types and of semantic relations. Semantic types are mostly broad subject matter

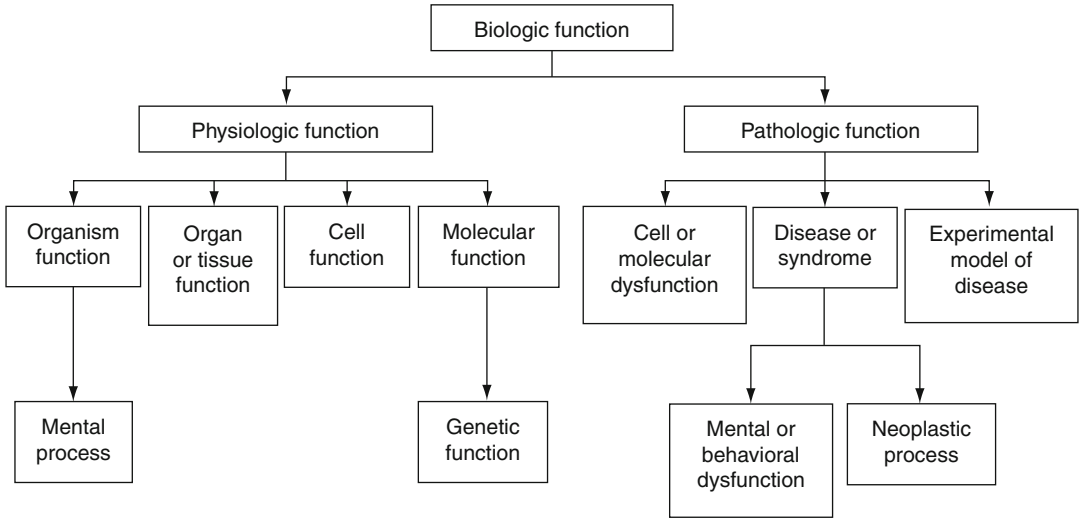


Fig. 8.4 UMLS semantic type examples

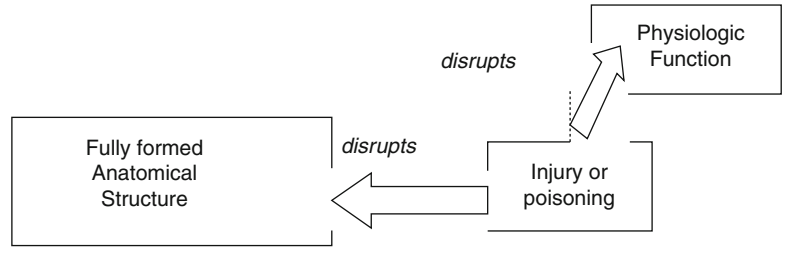


Fig. 8.5 Semantic relationships between semantic types of the UMLS

categorizations, like disorder or syndrome or clinical drug. Semantic relationships are useful linkages that exist between different semantic types. For example, clinical drug *treats* disease or syndrome. The Semantic Network can be used in computer systems to help interpret meaning or to trigger a clinical decision support rule.

For example, some of the semantic types are shown in Fig. 8.4.

The Semantic Network consists of:

- Semantic types (high-level categories)
- Semantic relationships (relationships between and among semantic types)

The Semantic Network is used to categorize any medical vocabulary that contributes to the UMLS.

There are 133 semantic types in the Semantic Network. Every Metathesaurus concept is assigned at least one semantic type. Some concepts have been assigned as many as five semantic types. Semantic types are provided in the Metathesaurus file named MRSTY.RRF.

Table 8.3 Relationships between semantic types

Semantic type	Semantic relationship	Semantic type
Injury or poisoning	Disrupts	Fully formed anatomical Structure
Injury or poisoning	Disrupts	Physiologic function

Semantic types and semantic relationships together create a Semantic Network that represents the information relevant to the domain of biomedicine.

Semantic types and relationships can help systems, developers, and users with their efforts to interpret the meaning of a Metathesaurus concept.

Figure 8.5 illustrates two semantic relationships between semantic types (Table 8.3).

The use of the semantic relations is directional. The object and specification of the relations are not necessary reciprocal in nature, as demonstrated in these examples.

The information associated with each semantic type includes:

- A unique identifier
- A tree number indicating its position in the "Isa" hierarchy
- A definition (human readable definition)
- Its immediate parent and immediate children

The information associated with each relationship includes:

- A unique identifier
- The semantic type of the relationship
- A tree number
- A definition (human readable)
- Examples (for most relations)
- The set of semantic types that can be expected to be linked together using this relationship

An example Semantic Network record is:

STY|T020|Acquired Abnormality|A1. 2. 2. 2|An abnormal structure, or one that is abnormal in size or location, found in or deriving from a previously normal structure. Acquired abnormalities are distinguished from diseases even though they may result in pathological functioning (e.g. "hernias incarcerate"). |Abscess of prostate; Hemorrhoids; Hernia, Femoral; Varicose Veins||||| Acquired Abnormality|co-occurs_with|Injury or Poisoning|D| Acquired Abnormality|isa|Anatomical Abnormality|D| Acquired Abnormality|result_of|Behavior|D|

Examples of the semantic types used in the network are:

- Organisms
- Anatomical structures
- Biologic function
- Chemicals
- Events
- Physical objects
- Concepts or ideas

Semantic types are classified into a hierarchy which is organized under two main categories: **Entity** and **Event**.

Examples of **Entity** semantic types are:	Examples of **Event** semantic types are:
Amphibian	Social behavior
Gene or genome	Laboratory procedure
Carbohydrate	Mental process

Semantic types exist in various levels of granularity (i.e., specificity). In the UMLS Metathesaurus, semantic type is assigned at the most specific level possible based on the information available. For example, the concept "metatarsal" would receive the semantic type "bone of the foot," not the semantic type "bone of the lower extremity" because "bone of the foot" is a more specific concept.

Although there are 54 semantic relations, the most common link between most semantic types is the "Isa" relationship. The "Isa" relationship establishes the hierarchy of types within the Semantic Network and is used as the basis for deciding which type is the most specific semantic type available for assignment to any specific UMLS Metathesaurus concept.

Some examples of the "Isa" relationships are:

- Animal Isa entity
- Carbohydrate Isa chemical
- Human Isa mammal

There are five major, nonhierarchical relationships which are:

- Physically related to
- Spatially related to
- Temporally related to
- Functionally related to
- Conceptually related to

Semantic relationships assigned at higher levels in the hierarchy may or may not hold at the concept level.

For example, the relationship "clinical drug treats disease or syndrome" does not hold at the concept level for *Tylenol* and *malignant neoplasm*. Tylenol does not treat malignant neoplasms.

Not all relationships that apply at the concept level are indicated in the Semantic Network. For example, the fact that one third of type I diabetics will go on to develop renal insufficiency does not imply that disorders cause renal insufficiency.

One of the more important relationships within the Semantic Network is the Isa, parent–child, or broader–narrower, relationship. This relationship illustrates the hierarchies that exist between biomedical types of concepts and relations. Child (narrower) relationships can be thought of as a specific form of a "subtype." For example, the semantic type "biologic function" is the parent of, or broader than, the semantic type "physiologic function."

Table 8.4 Examples of parent–child relationships

Parent (broader) type	Child (narrower) type
Physiologic function	Organism function
Affects	Disrupts
Finding	Sign or symptom

The level of granularity varies across the UMLS Semantic Network. For example, manufactured object is a child of physical object. Manufactured object has only two child concepts: medical device and research device. It is true that there are manufactured objects other than medical devices or research devices. Rather than expand the number of semantic types, concepts that are neither medical devices nor research devices have been simply assigned the broader semantic type manufactured object. This is problematic due to the creation of modeling inconsistencies. The compromise was made to keep the Semantic Network from growing to become too complex to implement.

Some examples of the parent–child relationships are shown in Table 8.4.

SPECIALIST Lexicon and Lexical Tools

The SPECIALIST Lexicon is an English language lexicon containing many words commonly found in the biomedical domain. Words in the lexicon were selected from a variety of sources including MEDLINE® abstracts, *Dorland's Illustrated Medical Dictionary*, and the general English vocabulary. The majority of the words are nouns and noun phrases.

The lexicon is made up of a set of lexical entries. Each entry represents a word (called a lexical item). The entry can provide one or more spellings, for example, in a particular part of speech, and describes the morphological, orthographic, and syntactic properties of the word.

The lexical tools are a collection of Java programs that process natural language words and terms including free text narratives. The lexical tools include a word normalize (i.e., Norm), a word index generator, and a lexical variant generator (LVG). When used together, the SPECIALIST Lexicon and Lexical Tools provide users with a head start toward developing and using natural language processing systems.

The SPECIALIST Lexicon is an English language lexicon (dictionary) which includes biomedical domain specific terms as well as commonly occurring English words and phrases. The lexical entry for each word or term (noun phrase) provides the following information:

- Syntactic information
- Morphological information (inflection, derivation, and composition)
- Orthographic information (spelling)

Currently the SPECIALIST Lexicon contains over 200,000 terms and is used as input to the UMLS lexical tools as an aid for natural language processing. Terms are selected for inclusion in the SPECIALIST Lexicon from a variety of sources including:

- The UMLS Test Collection of MEDLINE abstracts
- *Dorland's Illustrated Medical Dictionary*
- *The American Heritage Word Frequency Book*
- *Longman's Dictionary of Contemporary English*
- Current MEDLINE citation records

There is one UMLS SPECIALIST Lexicon lexical entry for each spelling or set of spelling variants appropriate for use in a particular part of speech. Each lexical record contains information regarding:

- Base form of the term
- Its part of speech
- Its unique identifier
- Its rules for spelling variants

The base form of a term is the uninflected form of the original term; this would be the singular form in the case of a noun, and would be the infinitive form in the case of a verb, and would be the positive form in the case of an adjective or adverb.

The UMLS SPECIALIST Lexicon recognizes eleven parts of speech which are:

Verbs	Pronouns
Nouns	Prepositions
Adjectives	Conjunctions
Adverbs	Complementizers
Auxiliaries	Determiners
Modals	

Approximately 86% of the entries are nouns or noun phrases.

The UMLS SPECIALIST Lexicon's lexical tools are a set of Java programs designed to aid in the processing of natural language.

The three most commonly used Java programs are:

- The lexical variant generator (LVG)
- The normalized string generator (Norm)
- The word index generator (Wordind)

Together these programs help to address the high degree of variability in healthcare natural language. Words often have several inflected forms that are properly considered instances of the same word with respect to meaning. The verb "treat," for example, has three inflectional forms: "treats," the third person singular present tense form; "treated," the past and past participle form; and "treating," the present participle form. Multiword terms in the Metathesaurus and other controlled vocabularies may have word order variants in addition to their variants in inflection or alphabetic and case-based variants. These lexical tools allow the user to aggregate this sort of variation. The goal is to aggregate terms to a meaning as represented by a concept (in the UMLS, this would be identified by a CUI).

The lexical variant generator contains a series of commands that can be chosen to perform lexical transformations from text. These commands handle lexical variations such as:

- Inflections and conjugations
- Word order in multiword terms
- Alphabetic case
- Punctuation
- Possessives

Developers can create their own sequence of commands to process text in a way that meets the needs of their applications and users. The Norm program is a predefined set of commands from the LVG program packaged together designed to produce a normalized form of a word.

The lexical tool's Java system, **Norm,** is used to create the normalized strings for terms included in the SPECIALIST Lexicon. The normalization process involves removing any possessives, replacing all punctuation with spaces, and removing stop words (words too common to practically index or of no informational value) such as "No Other Specification" or NOS; the program reduces all words to lower case; it breaks strings into their constituent words and sorts its words so that they appear in alphabetic order.

Table 8.5 Example of the NLP handling of the phrase Hodgkin's disease, NOS

	Hodgkin's diseases, NOS
Remove genitive	Hodgkin diseases, NOS
Remove stop words	Hodgkin diseases,
Lowercase	hodgkin diseases,
Strip punctuation	hodgkin diseases
Uninflect	hodgkin disease
Sort words	disease hodgkin

Table 8.5 is an example of the sequential steps of Norm's normalization process for the term "Hodgkin's diseases, NOS."

The Norm program has been employed in actual systems to:

- Find similar terms
- Map terms to UMLS concepts
- Find lexical variants for an input term often from natural language

Wordind creates word indices by separating an input string into a unique list of lowercased "words." Stop words are removed from the output. Wordind defines a word as a sequence of one or more alphanumeric characters.

For example, the phrase "Increased heart rate in an overweight forty-year-old male" would generate:

- increased
- heart
- rate
- overweight
- forty
- year
- old
- male

Wordind reads from structured input and writes to standard output. It outputs one line per non-stop word. This tool is used by the UMLS to produce, MRXW.RRF, the word index for the Metathesaurus.

NCI Enterprise Vocabulary Services

National Cancer Institute's (NCI) Enterprise Vocabulary Services (EVS) provides resources and services to meet the NCI's needs for controlled terminology and to facilitate the standardization of terminology and information systems across the institute and also the biomedical research community.

EVS terminological resources include the:

- NCI Thesaurus is a compendium of mapped terminological resources used in a growing number of NCI systems. It provides rich textual and ontological descriptions of 80,000 central biomedical concepts.
- NCI Metathesaurus is a biomedical terminological database, connecting 3,600,000 terms from more than 70 terminologies.
- NCI Term Browser publishes all terminologies hosted by NCI EVS in an integrated environment, providing search, cross-links, and a user-friendly interface to ICD9-CM, CTCAE, MedDRA, SNOMED CT, NDF-RT, GO, and many other terminologies and ontologies used by NCI and its partners.

These and other resources and services are described on the NCI EVS website, and on the companion EVS NCI Wiki and EVS caBIG websites.

EVS is a service of the NCI's Center for Biomedical Informatics and Information Technology (CBIIT). It is a key component of the cancer Common Ontological Resource Environment (caCORE) and the cancer Biomedical Informatics Grid (caBIG).

WHO Family of Classifications

ICD9-CM (as a Specific Version of the International Classification of Diseases)

ICD is part of the World Health Organization (WHO) family of classifications. Committees are formed with experts from member nations and are made available for international use. The WHO publishes these international classifications for health so that there will be a consensual, meaningful, and useful framework which governments, providers, and consumers can use as a common language.

Internationally endorsed classifications are meant to facilitate the storage, retrieval, analysis, and interpretation of data. They also are designed to allow the comparison of data within populations over time and between populations as well as the compilation of nationally consistent data.

The purpose of the WHO Family of International Classifications is to promote the appropriate selection of classifications in the range of global settings in health and healthcare.

Sir George Knibbs, from Australia, cited François Bossier de Lacroix (1706–1777), better known as Sauvages, with the first attempt to classify diseases systematically (10). *Nosologia Methodica, a* comprehensive treatise, was published by Sauvages. Linnaeus (1707–1778) authored *Genera Morborum*. At the beginning of the nineteenth century, the classification of disease in widest use was published by William Cullen (1710–1790), from Edinburgh, which was published in 1785 and was entitled *Synopsis Nosologiae Methodicae*.

The statistical study of disease began a century earlier based upon the work of John Graunt in creating the London Bills of Mortality. The kind of classification envisioned by Graunt is exemplified by his attempt to estimate the proportion of live-born children who died before reaching the age of 6 years; no records of age at death were available. He took all deaths classified as either due to thrush, convulsions, rickets, teeth and worms, abortives, chrysomes, infants, liver-grown, and added to them half the deaths classified as either smallpox, swinepox, measles, and due to worms without convulsions.

Despite the crudity of this measure, his estimate of 36% mortality before the age of 6 years appears from later evidence to have been correct. While three centuries have contributed substantively to the scientific accuracy of disease classification, there are still difficulties related to the classification of diseases and causes of death.

Fortunately for the progress of preventive medicine, the General Register Office of England and Wales, at its inception in 1837, appointed William Farr (1807–1883) – as its first medical statistician – a man who not only made the best possible use of the imperfect existing classifications of disease available at the time but also worked toward improved classifications and toward greater international uniformity in their application. Farr found that the classification of Cullen did not embody the advances of medical science, nor was it satisfactory for statistical purposes. In the first Annual Report of the Registrar General, therefore, he discussed the principles that should govern a statistical classification of disease and urged the adoption of a uniform classification as follows:

> The advantages of a uniform statistical nomenclature, however imperfect, are so obvious, that it is surprising no attention has been paid to its enforcement in Bills of Mortality. Each disease has, in many instances, been denoted by three or four terms, and each term has been applied to as many different diseases: vague, inconvenient names have been employed, or complications have been registered instead of primary diseases. The nomenclature is of as much importance in this department of inquiry as weights and measures in the physical sciences, and should be settled without delay.

The nomenclature and statistical classification received constant consideration by Farr in his annual "Letters" to the Registrar General published in the Annual Reports of the Registrar General. The utility of a uniform classification of causes of death was uniformly praised at the first International Statistical Congress, held in Brussels in 1853, that the Congress requested that William Farr and Marc d'Espine of Geneva prepare an internationally applicable and uniform classification of causes of death. At the next Congress, in Paris in 1855, Farr and d'Espine submitted two separate terminologies which were based on very different organizational paradigms. Farr's classification was organized into five hierarchies: epidemic diseases, constitutional (general) diseases, local diseases arranged according to anatomical site, developmental diseases, and diseases that are the direct result of violence.

The congress adopted a terminology consisting of 139 terms. In 1864, this classification was revised in Paris based upon Farr's model and was subsequently further revised in 1874, 1880, and 1886. Although this classification was never universally adopted, the general principles proposed by Farr, including the method for classifying diseases by anatomical site, have survived as the conceptual basis for the International List of Causes of Death.

The International Statistical Institute, which followed the International Statistical Congress, during their 1891 Vienna meeting charged a committee, chaired by Jacques Bertillon (1851–1922), who was the Chief of Statistical Services of the city of Paris, with the preparation of a classification of causes of death. Bertillon, the grandson of Achille Guillard, was a noted botanist and statistician that had introduced the resolution requesting Farr and d'Espine to create a uniform classification at the 1853 first International Statistical Congress. The report of this committee was presented by Bertillon at the 1893 meeting of the International Statistical Institute in Chicago. The classification proposed by Bertillon was based on the classification of causes of death in use by the city of Paris at that time, which represented a synthesis of the English, German, and Swiss classifications. The classification was based on the principle of separately classifying general diseases and those localized to a particular organ or at a particular anatomical site.

The Bertillon Classification of Causes of Death received general approval and was adopted by several countries and many cities. The classification was first employed in North America by Jesús E. Monjarás for the statistics of San Luis de Potosí, Mexico. The American Public Health Association (APHA), in 1898, recommended the adoption of the Bertillon Classification as a health statistic in Canada, Mexico, and the United States. The APHA suggested that the classification should be revised every 10 years.

In 1945, the American Secretary of State created the United States Committee on Joint Causes of Death which was chaired by Lowell J. Reed, a Professor of Biostatistics at Johns Hopkins University. This committee included

representatives from the governments of Canada and the United Kingdom and from the Health Section of the League of Nations. The committee thought it would be advantageous to consider classifications from the point of view of morbidity and mortality. The committee recommended that the "various national lists in use should, as far as possible, be brought into line with the detailed International List of Causes of Death." The committee noted that the classification of sickness and injury is closely related to the classification of causes of death. The view that such lists are fundamentally different arose from the erroneous notion that the International List was a classification of terminal causes of death, whereas it was, in fact, based upon the disorder that initiated the train of events that ultimately resulted in the patient's death. The committee believed that not only should the classification of diseases for both morbidity and mortality statistics be comparable, but also there should be a single classification. At this time, an increasing number of statistical organizations were keeping health records involving both illness and death. A single classification would greatly facilitate health statistics' coding operations. A combined classification would also provide a common base for comparison of morbidity and mortality statistics.

A subcommittee then prepared a draft Statistical Classification of Diseases, Injuries, and Causes of Death and was after testing adopted by Canada, the United States, and the United Kingdom.

The WHO International Conference for the ninth revision of the International Classification of Diseases was held in Geneva on September 30, 1975. There had been a tremendous growth of interest in employing the ICD. A number of presentations were made by specialty organizations that had become interested in using the ICD for their own statistical purposes. Some specialty areas within the classification were under considerable pressure to provide more detail and to classify conditions by the part of the body affected rather than to those chapters dealing with the underlying generalized disease process. Other countries were less interested in a detailed and sophisticated classification, but needed a classification based on the ICD in order to assess their public health progress and how well they were controlling outbreaks of disease.

The final categorization accepted by the conference retained the basic structure of the ICD, but added additional detail specifically at the level of the four-digit subcategories, and also added optional five-digit subdivisions.

Recognizing the needs of countries not requiring this level of detail, the design required that the three-digit level categories were well formed. The ninth revision includes an optional alternative method of classifying diagnostic statements that included information about an underlying general disease and a manifestation of the disease in a particular organ or body site. This classification system is known as the dagger and asterisk system and has been retained in the tenth revision of the ICD.

The present coding practices employed in the creation of the ICD series rely on data methods and principles for terminology maintenance that have changed little since the adoption of the statistical bills of mortality in the mid-seventeenth century [2]. The most widely accepted standard for representing patient conditions, ICD9-CM [3], is an intellectual descendent of this tradition. ICD9-CM relies overwhelmingly on a tabular data structure with limited concept hierarchies and no explicit mechanism for synonymy, value restrictions, inheritance, or semantic and nonsemantic linkages. The Center for Disease Control and Prevention's national committee for health and vital statistics for many years had the responsibility of creating, maintaining, and distributing the clinical modification of the ICD for the United States (ICD9-CM). They have more recently changed their name to the National Committee on Health Statistics. The maintenance environment for this healthcare classification is a word processor, and its distribution is nearly exclusively paper-based.

In the United States, ICD9-CM is the standard for representing morbidity in patient populations and is used for billing in the United States. Diagnostic-related groups which are used for capitated billing in the hospital setting are based on the ICD9-CM codes selected for the patient

based on their clinical condition. Mortality coding in the United States is performed by assigning ICD10 codes.

Similar limitations exist in the maintenance environment of ICD10 – the tenth revision of the International Statistical Classification of Diseases and Related Health Problems, which is being adopted as the national standard for diagnosis coding in an increasing number of countries. The WHO is working on its eleventh revision.

Table 8.6 lists file names and sizes when extracted for the ICD9-CM files available by FTP, for use as of October 1, 2010, for federal fiscal year 2011 (FY11). Please note that Appendix B, the Glossary of Mental Disorders, has been retired and is no longer distributed with the ICD9-CM.

A copy of the table of contents for the coding guidelines follows. This shows the various chapters and gives the reader a hint as to the complexity of the coding rules. The classification has many parts including the primary terms and entry terms which are effectively synonyms. The terminology contains procedures as well as diagnoses and lists of drugs and neoplasms and accidents. Although ICD is most commonly known as a diagnostic terminology, ICD10-PCS has a robust procedure coding system (Table 8.7).

ICD9-CM/ICD10-CM

This classification relating to clinical modification is developed in the USA. For more details of both classifications, please contact: WHO Collaborating Center for the Classification of Diseases for North America, Data Policy and Standards Staff, Office of the Center Director, National Center for Health Statistics, Centers for Disease Control and Prevention (CDC), Room 1100, 6525 Belcrest Road, Hyattsville, MD 20782, USA.

ICD10-AM

Australian Modification of ICD10 (in preparation) scheduled to come into effect in Australia and New Zealand on July 1, 1998. Contact: WHO Collaborating Centre for the Classification of Diseases, Australian Institute of Health and Welfare, GPO Box 570, Canberra ACT 2601, Australia.

ICPC

International Classification of Primary Care published by WONCA (World Organization of Family Doctors).

Table 8.6 ICD files and their size and contents

Filename	Size in bytes	Expanded files	Expanded file sizes	Description
APPNDX11.ZIP	80,500	DMORPH11.RTF	183,309	Appendix A, Morphology of Neoplasms
		DDRGCL11.RTF	83,011	Appendix C, Classification of Drugs
		DINDST11.RTF	64,048	Appendix D, Classification of Industrial Accidents
		DC_3D11.RTF	260,850	Appendix E, List of Three-Digit Categories
DDRUGS11.ZIP	91,663	DDRUGS11.RTF	979,633	Table of Drugs and Chemicals
DINDEX11.ZIP	1,139,425	DINDEX11.RTF	18,507,209	Index to Diseases
EINDEX11.ZIP	71,986	EINDEX11.RTF	1,186,318	Index to External Causes
DTAB11.ZIP	755,064	DTAB11.RTF	13,012,446	Tabular List of Diseases
PINDEX11.ZIP	279,074	PINDEX11.RTF	3,860,698	Index to Procedures
PTAB11.ZIP	235,984	PTAB11.RTF	3,976,227	Tabular List of Procedures
PREFAC11.RTF	73,691	---	---	Preface

Table 8.7 ICD Table of Contents to show the level of specification to the coding rules

(continued)

Table 8.7 (continued)

Table 8.7 (continued)

(continued)

Table 8.7 (continued)

Contact: Chairman, WONCA International Classification Committee, Institute of Community Health, Department of General Practice, Winslowsparken 17, DK-5000 Odense C, Denmark.

ATC Classification

Anatomical Therapeutic Chemical classification system for drugs. Contact: WHO Collaborating Centre for Drug Statistics Methodology, P.O. Box 100, Veitvet, N-0518 Oslo 5, Norway.

WHOART

WHO Adverse Reactions Terminology

Contact: WHO Collaborating Centre for International Drug Monitoring, P.O. Box 26, S-751 03 Uppsala, Sweden.

Dermatology

The International League of Dermatological Societies has been granted permission by WHO to prepare an application of ICD10 to dermatology. The work is being carried out by the British Association of Dermatologists.

Pediatrics

An application of ICD10 to pediatrics is being prepared by the Royal College of Paediatrics and Child Health under the auspices of the International Pediatric Association in accordance with a copyright agreement with WHO. Contact: Health Services Information Officer, Royal College of Paediatrics and Child Health, 50 Hallam Street, London W1N 6DE, England.

Rheumatology and Orthopedics

WHO has granted permission for the International League of Associations for Rheumatology (ILAR) to prepare an application of ICD10 to rheumatology and orthopedics.

Below is an example of the format of the ICD9 codes in the field of neuropsychiatric disorders (Tables 8.8 and 8.9):

ICD9-CM breaks some good practices of terminology development. The identifiers are meaningful, and therefore if you need to insert a term in between two others, you may not have space in the identifier structure. ICD9-CM does not have formal definitions and cannot be used compositionally. The terminology was developed with the use case of healthcare billing and administration as the strongest director of its development decision making. Therefore, where clinical concerns are at odds with its main purpose, they have not been well served. That said, ICD9-CM has been used in more clinical research studies than any other terminology.

The error rate in assignment of ICD9 codes ranges from 23% to 40% in the health informatics literature. Most of these errors in assignment are in the level of granularity (location within the hierarchy) rather than miss-coding across hierarchies.

ICD10 is used in our country for mortality coding and is about four times larger than ICD9 in the number of terms. ICD10-PCS has contributed to the International Classification of Healthcare Interventions (ICHI) which extends

Table 8.8 ICD9 codes for defining neuropsychiatric events

ICD9	Description	Subcategory	Category
290.40	Vascular dementia, uncomplicated	Dementia	Psychiatric event
290.41	Vascular dementia with delirium	Dementia	Psychiatric event
290.42	Vascular dementia with delusion	Dementia	Psychiatric event
290.43	Vascular dementia with depressed mood	Dementia	Psychiatric event
290.8x	Other specified senile psychotic conditions	Psychoses	Psychiatric event
290.9x	Unspecified senile psychotic condition	Psychoses	Psychiatric event
292.1x	Drug-induced psychotic disorders	Drug induced	Psychiatric event
292.2x	Pathological drug intoxication	Drug induced	Psychiatric event
292.8x	Other specified drug-induced mental disorders	Drug induced	Psychiatric event
292.9	Unspecified drug-induced mental disorder	Drug induced	Psychiatric event
293.xx	Transient organic psychotic conditions (delirium, delusion, psychosis)	Psychoses	Psychiatric event
294.xx	Persistent mental disorders due to conditions classified elsewhere	Psychoses	Psychiatric event
296.xx	Episode mood disorders	Psychoses	Psychiatric event
297.xx	Delusional disorders	Psychoses	Psychiatric event
298.xx	Other nonorganic psychoses	Psychoses	Psychiatric event
300.0x	Anxiety states	Anxiety, stress, or depressive	Psychiatric event
308.xx	Acute reaction to stress	Anxiety, stress, or depressive	Psychiatric event
309.xx	Adjustment reaction, includes adjustment disorders	Anxiety, stress, or depressive	Psychiatric event
311.xx	Depressive disorder, not elsewhere classified	Anxiety, stress, or depressive	Psychiatric event
312.0x	Disturbance of conduct, not elsewhere classified	Anxiety, stress, or depressive	Psychiatric event

(continued)

Table 8.8 (continued)

ICD9	Description	Subcategory	Category
312.35	Disturbance of conduct, isolated explosive disorder	Anxiety, stress, or depressive	Psychiatric event
313.0x	Disturbance of emotions specific to childhood and adolescence	Anxiety, stress, or depressive	Psychiatric event
E950	Suicide and self-inflicted injury – poisoning by solid or liquid substances	Suicide – other	Psychiatric event
E951	Suicide and self-inflicted injury – poisoning by gases in domestic use	Suicide – other	Psychiatric event
E952	Suicide and self-inflicted injury – poisoning by other gases and vapors	Suicide – other	Psychiatric event
E953	Suicide and self-inflicted injury – by hanging, strangulation, and suffocation	Suicide – other	Psychiatric event
E954	Suicide and self-inflicted injury – by drowning	Suicide – other	Psychiatric event
E955	Suicide and self-inflicted injury – by firearms, air guns, and explosives	Suicide – other	Psychiatric event
E956	Suicide and self-inflicted injury – by cutting and piercing instrument	Suicide – cutting	Psychiatric event
E957	Suicide and self-inflicted injury – by jumping from high place	Suicide – other	Psychiatric event
E958	Suicide and self-inflicted injury – by other and unspecified means	Suicide – other	Psychiatric event
E959	Suicide and self-inflicted injury – late effects of self-inflected injury	Suicide – other	Psychiatric event
323.4x	Other encephalitis due to infection classified elsewhere		Encephalitis
323.5x	Encephalitis following immunization procedure		Encephalitis
323.6x	Postinfectious encephalitis		Encephalitis
323.7x	Toxic encephalitis		Encephalitis
323.8x	Other causes of encephalitis		Encephalitis
323.9x	Unspecified cause of encephalitis		Encephalitis
348.3x	Encephalopathy, not elsewhere classified		Encephalitis
327.0x	Organic insomnia		Disturbances of consciousness
780.0x	Alteration of consciousness		Disturbances of consciousness
780.1x	Hallucinations		Disturbances of consciousness
780.2x	Syncope and collapse		Disturbances of consciousness
780.4x	Dizziness and giddiness		Disturbances of consciousness
780.54	Other hypersomnia		Disturbances of consciousness
781.0x	Abnormal involuntary movements		Abnormal movements
345.xx	Epilepsy		Seizure
780.3x	Convulsion		Seizure
430.xx	Subarachnoid hemorrhage		Stroke
431.xx	Intracerebral hemorrhage		Stroke

Table 8.8 (continued)

ICD9	Description	Subcategory	Category
432.xx	Other and unspecified intracranial hemorrhage		Stroke
433.xx	Occlusion and stenosis of precerebral arteries		Stroke
434.xx	Occlusion of cerebral arteries		Stroke
435.xx	Transient cerebral ischemia		Stroke
436.xx	Acute, but ill-defined, cerebrovascular disease		Stroke
437.xx	Other and ill-defined cerebrovascular disease		Stroke
438.xx	Late effects of cerebrovascular disease		Stroke
368.1x	Subjective visual disturbance, unspecified		Vision disturbances
368.2x	Double vision		Vision disturbances
368.4x	Visual field defects		Vision disturbances
368.55	Acquired color vision deficiencies		Vision disturbances
368.59	Other color vision deficiencies		Vision disturbances
368.6x	Night blindness		Vision disturbances
368.9x	Unspecified visual disorders		Vision disturbances
780.52	Insomnia NOS		Other neuro events
781.1x	Taste disturbance		Other neuro events
781.99	Other symptoms involving nervous and musculoskeletal systems		Other neuro events
784.0x	Headache		Other neuro events

Table 8.9 ICD9 codes for defining suicide and accidents

ICD9	Description	Subcategory	Category
E880–E888	Fall	Fall	Accident
E810–E829	Motor vehicle accident – specified (i.e., on highway etc.)	Vehicle	Accident
E846–E848	Motor vehicle accident – not specified	Vehicle	Accident
E850–E869	Accidental poisoning by drugs, medical substances, biologicals, liquid, gases, and vapors	Poisoning	Accident
E870–E879	Accident or reaction caused by medical procedures or treatments	Other	Accident
E890–E999	Accidents caused by fire and flames	Other	Accident
E830–E845	Water transport accident	Other	Accident
E800–E807	Railway accidents	Other	Accident
E950	Suicide and self-inflicted injury – poisoning by solid or liquid substances	Suicide – other	Suicide
E951	Suicide and self-inflicted injury – poisoning by gases in domestic use	Suicide – other	Suicide
E952	Suicide and self-inflicted injury – poisoning by other gases and vapors	Suicide – other	Suicide
E953	Suicide and self-inflicted injury – by hanging, strangulation, and suffocation	Suicide – other	Suicide

(continued)

Table 8.9 (continued)

ICD9	Description	Subcategory	Category
E954	Suicide and self-inflicted injury – by drowning	Suicide – other	Suicide
E955	Suicide and self-inflicted injury – by firearms, air guns, and explosives	Suicide – other	Suicide
E956	Suicide and self-inflicted injury – by cutting and piercing instrument	Suicide – cutting	Suicide
E957	Suicide and self-inflicted injury – by jumping from high place	Suicide – other	Suicide
E958	Suicide and self-inflicted injury – by other and unspecified means	Suicide – other	Suicide
E959	Suicide and self-inflicted injury – late effects of self-inflected injury	Suicide – other	Suicide

PCS to include a wide variety of healthcare interventions. The International Classification of Function (ICF) is a terminology which represents health status outcomes for statistical coding purposes. This is an important step in formalizing the information used for health outcomes studies. These terminologies can be obtained from the WHO website at http://www.who.int/classifications/en/ and are available in multiple languages.

Current Procedural Terminology® (CPT®)

Scope and Purpose

The American Medical Association developed and maintains the Current Procedural Terminology® (CPT®) code set. In the United States, CPT is virtually used by all public and private healthcare payers, all healthcare professionals, and institutional providers.[1] Primarily used to report services and procedures reported on health insurance claims, the first edition of CPT was published in 1966.[2] The current edition includes numerical codes with descriptors for reporting medical services and procedures performed by physicians and other healthcare professionals. In the context of reporting services, CPT provides a consistent language to describe medical, surgical, and diagnostic services.[1]

History of CPT

In the past five decades, the CPT code set has evolved significantly. The first edition was designed to promote the use of standard terms and descriptors to document procedures in medical records and contributed basic information for insurance risk evaluation and statistical analysis. The first edition used a four-digit numbering system and primarily included surgical procedures. In 1970, the second edition introduced a five-digit numbering system and additional terms and codes to describe procedures and services. The third and fourth editions introduced the CPT Editorial Panel to guide development and updating of the code set by meeting three times a year.[3]

From the mid-1980s through 2000, the CPT code set experienced significant growth and adoption in federal programs. In 1983, it was adopted as Level I of the Healthcare Common Procedure Coding System (HCPCS) developed by the Centers for Medicare and Medicaid Services (CMS) requiring the use of HCPCS to report services for Part B of the Medicare

[1]CPT: History and Role in the U.S. Healthcare System, American Medical Association, 2004

[2]American Medical Association, *Current Procedural Terminology* (*CPT®*), 2011 Forward p. v

[3]American Medical Association, *Principles of CPT® Coding*, 6th edn., 2010, p. 2

program and, a few years later, CMS also required the use of HCPCS by state Medicaid agencies. Subsequently, as part of the Omnibus Budget Reconciliation Act, CPT codes were required for reporting outpatient hospital surgical procedures. In 1992, CPT was associated with the Resource-Based Relative Value Scale (RBRVS), Medicare Physician Fee Schedule to analyze the utilization of services. In 2000, the Department of Health and Human Services designated the CPT code set as the national coding standard for physician and other healthcare professional services and procedures under the Health Insurance Portability and Accountability Act (HIPAA). The HIPAA designation required use of the CPT code set for all financial and administrative healthcare transactions transmitted electronically.[3]

Structure of CPT

The CPT code set includes three categories (I, II, and III). This section discusses Category I CPT codes.

The CPT Category I terminology classifies procedures and services into six major sections that include Surgery, Anesthesiology, Radiology, Pathology and Laboratory, Medicine, and Evaluation and Management. Each section is organized according to traditionally recognized body systems and clinical practice.[4] For example, the Surgery section is subdivided into subsections by body system that includes Integumentary, Musculoskeletal, Respiratory, Cardiovascular, Urinary, Nervous, and Eye and Ocular Adnexa. The Radiology section includes subsections of Nuclear Medicine and Diagnostic Ultrasound.

Each service or procedure includes a descriptive term and an associated five-digit numerical code. Like some other classification systems, the CPT numerical code is "meaningful" where the first digit generally denotes the section that the procedure or service is assigned. For example, procedures or services classified in the Digestive System subsection will most likely have a code assignment that begins with the number 4 (e.g., CPT code 40490 *Biopsy of lip*[5]), and those procedures or services classified in the Nervous System subsection will have a code assignment that begins with the number 6 (e.g., CPT code 61000 *Subdural tap through fontanelle, or suture, infant, unilateral or bilateral; initial*[5]). It is important to note that the assignment of a procedure or service to a specific section does not restrict its use by a particular medical specialty.[6]

A meaningful numbering scheme is useful in that it facilitates visual searching, organization, and mental recall. However, like other classification systems that use meaningful numbering schemes, the CPT code set has experienced space limitations in some sections. Consequently, as of 2010, some sections of the CPT code set will include procedures or services that are conceptually consistent with other procedures or services, however may have code assignments that are not necessarily in numerical sequence.

Future of CPT

In 1998, the AMA convened the CPT-5 workgroup to examine methods for the CPT code set to adapt to the emerging needs of health information technology (HIT) and the expanding data requirements of physicians and healthcare professionals. Three major developments evolved from the workgroup recommendations:

1. Category II CPT codes were created to support the reporting of performance measure results using claims as a data source.
   Category II CPT codes are alphanumeric and include the letter F as the terminal character. They are developed to report the results from

[4]CPT: History and Role in the U.S. Healthcare System, American Medical Association, 2004

[5]CPT® © 2010 American Medical Association; All rights reserved

[6]American Medical Association, *Current Procedural Terminology* (*CPT®*), 2011, Forward p. X

the application of clinical quality measures that consider the strength of evidence, guideline recommendations, and gaps in care. Category II codes represent clinical findings and services where there is a robust evidence base for contributing to health outcomes and quality patient care. For example, Category II code 3016 F *Patient screened for unhealthy alcohol use using a systematic screening method*[5] is based on clinical guidelines from the United States Preventive Services Task Force[7]. The recommendations take into account evidence-based clinical practice guidelines and consensus standards. The CPT Editorial Panel and the Performance Measures Advisory Group develop and maintain the Category II codes.

2. Category III CPT codes were created to represent emerging technologies, procedures, and services.
   Similar to Category II codes, Category III codes are alphanumeric and include the letter T as the terminal character. They are temporary codes developed to support utilization, tracking of new procedures and services, and to support the approval process of the Food and Drug Administration (FDA).
3. A CPT data model was developed to address HIT requirements for computer readable information that can be interpreted with common meaning across healthcare systems.
   The CPT data model is intended to provide a structure to facilitate the integration of CPT data within electronic health records (EHRs) or other health information systems and support interoperability with other codes sets and terminologies. The data model provides a semantic structure, organizing CPT procedures and services into a hierarchy and assigning reference concepts. The reference concepts describe the essential characteristics of a procedure or service such as the anatomic site, device employed, substance used, or associated pathology. The semantic model does not alter the historical structure and payment focus of the CPT code set, yet it does provide technical elements to develop electronic files, software applications, and mappings to other code sets and terminologies. For example, software applications using the CPT data model will allow users to aggregate all CPT procedures on the knee using an arthroscope. Additionally, the reference concepts will facilitate mapping CPT arthroscopy procedures to similar procedures in other code sets.

[7]U.S. Preventive Services Task Force. Screening and behavioral counseling interventions in primary care to reduce alcohol misuse: recommendation statement. April 2004. Agency for Healthcare Research and Quality. Rockville, MD. Available at: http://www.ahrq.gov/clinic.3rduspstf/alcohol.alcomisrs.htm

LOINC [4]

Logical observation identifiers names and codes (LOINC) is a terminology which was initially developed to represent laboratory test names and then expanded to represent other clinical domains. It is maintained by the LOINC committee and has its origin at Regenstrief Institute of Indiana University. LOINC has its theoretical basis in Euclides which was a pathology-based representation formalism that originated in Europe.

The laboratory hierarchies of the LOINC terminology includes the usual categories of chemistry, hematology, serology, microbiology (including parasitology and virology), toxicology; as well as categories for drugs and the cell counts, and antibiotic susceptibilities. The clinical portion of the LOINC database includes entries for vital signs, hemodynamics, intake/output, EKG findings, obstetric ultrasound findings, cardiac echo findings, urologic imaging findings, gastroendoscopic procedures, pulmonary ventilator management, and selected survey instruments (e.g., Glasgow Coma Score, PHQ-9 depression scale, CMS-required patient assessment instruments).

LOINC is a precoordinated terminology. This means that LOINC codes have multiple components to their fully specified names. In LOINC, the fully specified name of a laboratory test result

or clinical observation has five or six main parts including: the name of the component or analyte measured (e.g., fasting blood sugar, or fosinopril), the property observed (e.g., the substance's concentration, mass, or volume), the timing of the measurement (e.g., is it measured over time or is it a momentary measurement), the type of sample (e.g., urine, serum), the scale of measurement (e.g., qualitative vs. quantitative), and where relevant, the method of the measurement (e.g., radioimmunoassay, immune blot). These can be described formally with the following syntax:

<Analyte/component>:<kind of property of observation or measurement>:<time aspect>:<system (sample)>:<scale>:<method>

The first part of the name can be further divided up into three subcomponents, separated by carats (^). The first subcomponent can contain multiple levels of increasing taxonomic specification, separated by dots (.). The third and fourth parts of the name (time aspect and system/sample) can also be modified by a second subpart, separated from the first by a carat. In the case of time aspect, the modifier can indicate that the observation is one selected on the basis of the named criterion (maximum, minimum, mean, etc.); in the case of system, the modifier identifies the origin of the specimen, if not the patient (e.g., blood donor, fetus, and blood product unit). The hierarchical structure is outlined in the list below.

Subpart name
Component/analyte
Name and modifier
Component/analyte name
Component/analyte subname
Component/analyte sub-subname
Information about the challenge (e.g., 1 H post 100 gm PO glucose challenge)
Adjustments/corrections
Kind of property (mass concentration, mass)

LOINC is available free of charge and can be downloaded along with a terminology browser RELMA from the LOINC website http://loinc.org/.

SNOMED CT

SNOMED itself was started in 1965 as SNOP (Systematized Nomenclature of Pathology) and then expanded into other medical fields. The College of American Pathologists (CAP) created the Systematized Nomenclature of Pathology (SNOP) and subsequently the Systemized Nomenclature of Medicine (SNOMED). In these systems, the number, scope, and size of the compositional structures have increased to the point where an astronomical number of terms can be synthesized from SNOMED atoms. One well-recognized limitation of this expressive power is the lack of syntactic grammar, compositional rules, and normalization of both the concepts and the semantics. Normalization is the process by which the system knows that two compositional constructs with the same meaning are indeed the same (e.g., that the term "colon cancer" is equivalent to the composition of "malignant neoplasm" and the site "large bowel"). These are issues addressed by CAP in their efforts to make SNOMED a robust reference terminology for healthcare [5, 6]. SNOMED developed into SNOMED II and then into SNOMED III which was also named SNOMED International. SNOMED then developed SNOMED RT which was the first description-logic-based large-scale healthcare clinical general medical terminology.

Other initiatives of importance are the Clinical Terms v3 (Read Codes), which are maintained and disseminated by the National Health Service (NHS) in the United Kingdom and the Galen effort, which expresses a very detailed formalism for term description. The Read Codes are a large corpus of terms, which is now in its third revision that is hierarchically designed and is used throughout Great Britain. Version one of the Read Codes has about 6,000 concepts, and version two had about 20,000 concepts. Version one was almost universally adopted in the UK. Version two had more variable uptake. During the time of the dissemination of version two, the UK began an effort to create their own large-scale clinical terminology Clinical Terms v3 which was a direct descendant of version two of the Read Codes. As of this writing we are saddened to learn of

Dr. James Read's untimely passing. The authors wish his family well and recognize his contribution to our field.

The Unified Medical Language System and its Metathesaurus have been developed in cooperation with the United States National Library of Medicine and provides an interlingua between many existing healthcare terminologies. It holds and distributes over 100 terminologies that are lexically mapped together and provided with a common concept unique identifier (CUI). The hierarchies of each terminology are maintained and not merged, although there is a Semantic Network that is assigned as types to the various concepts in the contributing terminologies.

A development of interesting note is the newly signed agreement of CAP and the NHS to merge the content of SNOMED RT and Clinical Terms Version 3 into a derivative work (announced 4/1999), which is named SNOMED Clinical Terms or SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms). This has been transferred to the International Healthcare Terminology Standards Development Organization (IHTSDO) which has its headquarters in Copenhagen, Denmark. In addition to English versions, there are Spanish, Danish, and French versions of SNOMED CT in various stages of development. IHTSDO website is found at www.ihtsdo.org.

SNOMED CT is a large general medical description-logic-based formal compositional terminology. Its polyhierarchical structure holds approximately 292,000 active concepts and over one million terms (i.e., preferred terms and synonyms). Other groups have added additional synonyms to SNOMED CT to improve its rate of capture of clinical text. There are specific rules and semantics covered by SNOMED CT that govern how one can use the content to create refined postcoordinated structures. By employing postcoordination, we can represent literally billions of concepts using SNOMED CT.

Only about 15% of SNOMED CT is fully defined and therefore can be autoclassified. This is a limitation of the power of the classifier over the range of the terminology. That said it is the largest logical terminology in healthcare and as such represents a significant step forward toward having available truly comparable and interoperable data from and to the practice of medicine.

SNOMED CT has been purchased by the US Federal Government and provided at no cost to any organization or individual with direct reporting responsibility to the US government. This includes most all healthcare organizations.

SNOMED CT aims to contribute to the improvement of patient care through underpinning the development of systems to accurately record healthcare encounters and to deliver decision support to healthcare providers. Ultimately, patients will benefit from the use of SNOMED CT to more clearly describe and accurately record their care, in building and facilitating better communication and interoperability in electronic health record exchange, and in creating systems that support healthcare decision making.

Benefits of SNOMED CT as Taken from the IHTSDO Website

General Benefits of SNOMED CT

- SNOMED CT is an international standard, has multilingual support, and enables the provision of a platform independent, cross-cultural, healthcare record.
- SNOMED CT provides a consistent terminology across all healthcare domains. This allows clinicians to communicate effectively and accurately across clinical domains and over the lifetime of a patient record.
- SNOMED CT allows precise recording of clinical information. By using many descriptions for a single clinical concept, it allows tailoring for individual care settings while maintaining consistency.
- SNOMED CT has an inherent structure. This provides for an unambiguous description of an individual concept in a logical way and allows application of logical processing and machine reasoning of clinical information.

- SNOMED CT can be extended in a controlled fashion to further enhance its usability and coverage.
- The recording of clinical data through SNOMED CT enables the consistent retrieval, transmission, and analysis of data from patient records across healthcare systems.
- SNOMED CT is well maintained and updated in collaboration with subject matter experts to represent current clinical knowledge

Operational Use

- SNOMED CT enables the capture of clinical information at a level of detail appropriate for the provision of healthcare.
- SNOMED CT enables patient data to be recorded by different people in different locations and to be combined into simple information views within the patient record. This enables the continuity of care across different care settings and locations.
- The consistent use of SNOMED CT reduces the risk of differing and incorrect interpretation of data in healthcare records by reducing the implicit contextual meaning associated with entered data.
- Appropriate use of SNOMED CT can contribute to the reduction of error rates and can help ensure the comprehensive recording of relevant data.
- Through sharing data, it can dramatically reduce the need to repeat health history at each new encounter with a healthcare professional.
- SNOMED CT enables efficient searching of patient records and retrieval of relevant clinical information.
- SNOMED CT facilitates point-of-care decision support, automatic identification of patient risk factors, and monitoring of response to treatment and adverse reactions to treatment. Using SNOMED CT to encode clinical information in the patient record, computers can assist the decisions made by healthcare professionals by providing contextually relevant information at the point of care, or by providing automated alerts, reminders, or checks.
- SNOMED CT can assist with identification of patients who match a given set of clinical criteria. For example, those who are eligible for a particular screening program, or a clinical trial, can be identified, and patients who are at a high risk of developing a given disease can be detected.
- SNOMED CT improves clinical efficiency by providing a standard clinically relevant terminology to the clinician for documentation of care.
- SNOMED CT's history mechanism enables clinical information collected over time to be meaningfully correlated together.

Secondary Use

- SNOMED CT can assist with public health monitoring. Encoding clinical information allows for the monitoring of diseases and disease trends at a population level. The more usable clinical information that is available, the easier it will be to tackle health issues or manage disease outbreaks.
- SNOMED CT enables the analysis of outcomes. There is an increasing focus on evidence-based medicine in clinical practice today, but little usable information to base that evidence on. Consistent use of SNOMED CT to code information in patient records will provide an improved information base to support outcome analysis.
- SNOMED CT can also facilitate performance analysis. As medicine moves toward evidence bases, fitness to practice and clinical revalidation are similarly moving toward performance-related measures. SNOMED CT can provide a consistent basis for evaluation.
- SNOMED CT enables the easier, more effective analysis of data.
- SNOMED CT will enable the provision of large populations of consistent data for medical research.
- SNOMED CT can facilitate process improvement activities by more consistent and accurate documentation of clinical events and activities and linking these to process measurements and timeliness of delivery of care.

Distribution of Information

- SNOMED CT can be used for the sharing and consistent distribution of outcome analysis data. For example, SNOMED CT can be used to analyze how many cancer surgeries are performed and to consistently record outcome data to determine whether surgery has an impact on long-term survival and local recurrence in cancer treatments. This type of outcome analysis will be invaluable once an evidence base is built up. If outcome data like this can be represented consistently using SNOMED CT, then a much wider pool of international data could be used to compare treatments both within and across countries, with resulting improvements in best practice.
- SNOMED CT can be used to setup and distribute decision support information in a consistent way.
- SNOMED CT can facilitate knowledge management through its standard terminology and the reference information embedded within. For example, SNOMED CT hierarchy can be used to aggregate similar kinds of information and knowledge together.

Updates and Maintenance of SNOMED CT

Development, Quality Assurance, and Release

SNOMED CT is continuously updated to meet the needs of users around the world. Revisions to the international version of SNOMED CT are released twice a year (see Fig. 8.6). Each release includes the core of the terminology (concepts, descriptions, and relationships), together with works to support the implementation and use of SNOMED CT, including subsets, cross maps to existing classifications and coding schemes, and an extensive set of guidelines.

These updates are driven by users of the terminology. Examples include refinements to descriptions, remodeling of concepts, or the addition of new concepts. Prior to release, the SNOMED CT content undergoes a clinical quality assurance process. A preliminary version is then prereleased to members for broader review before the final files are generated and distributed.

Improvements to SNOMED CT

Just as our knowledge about health and healthcare is constantly evolving, so too are healthcare terminologies which are living languages. As mentioned, the number of concepts in SNOMED CT is continuously growing. A continuous "cleaning up" process also takes place; from 2002 to 2008, approximately 20,000 concepts were deactivated because they were duplicates, outdated, ambiguous, etc. The quality of the descriptions is also improving. For example, more concepts now have sufficient logic definitions – particularly those in the disorder and procedure hierarchies.

You can assist your colleagues around the world in the ongoing development and

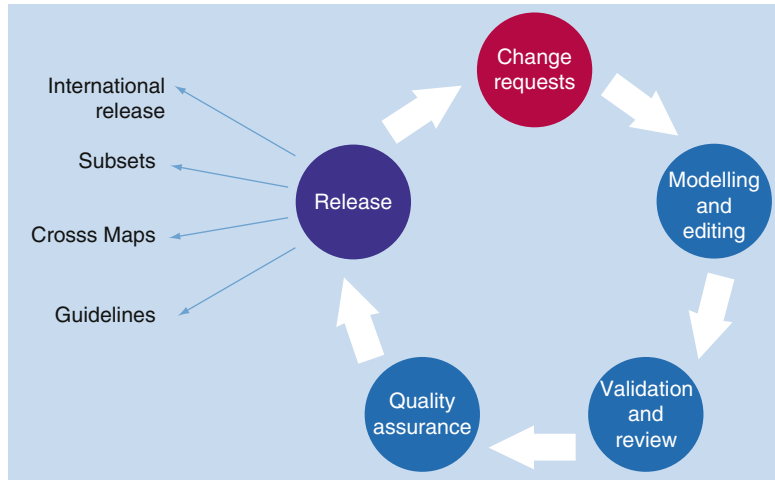


Fig. 8.6 SNOMED CT update schedule and dependencies

Fig. 8.7 SNOMED CT top level hierarchy

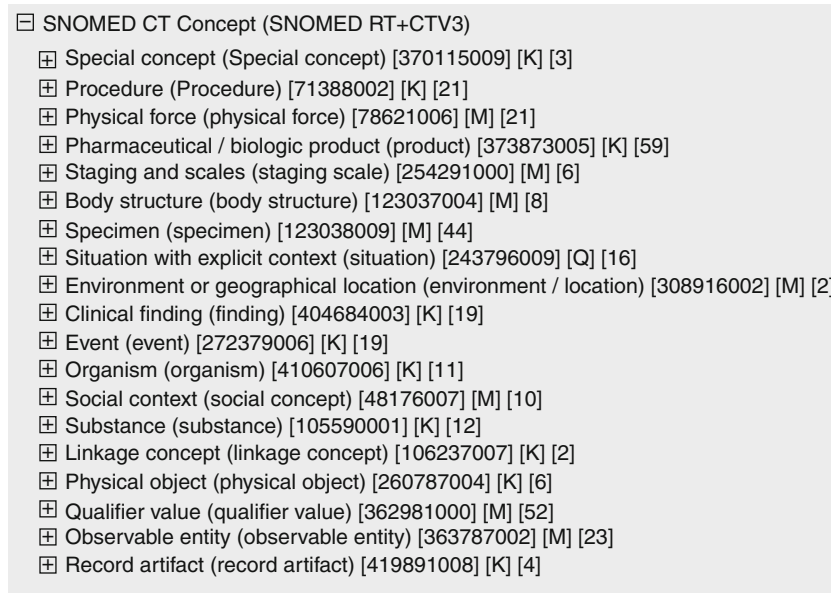
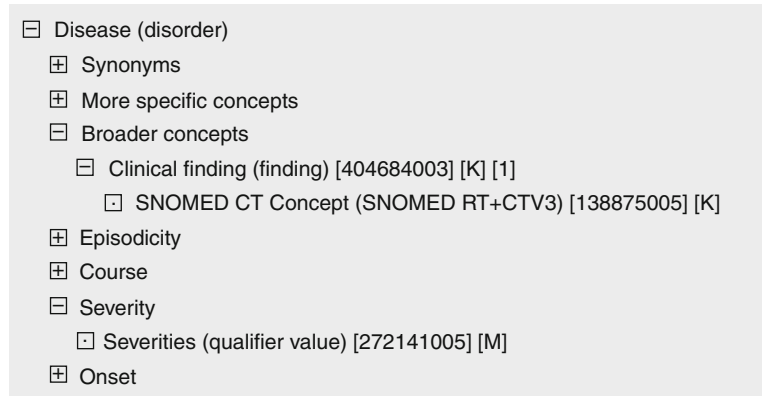


Fig. 8.8 SNOMED CT concept *disease* and its associated description logic definition



maintenance of SNOMED CT by becoming a member of IHTSDO's Working Groups and Special Interest Groups, by participating in discussions on the Collaborative Space and at face-to-face meetings, or by submitting requests for specific additions or changes to the standard. We also encourage you to share your experiences of using the standard through the Members' Operational Liaison Forum, the Affiliate Forum, and at local forums and events organized by IHTSDO's members.

This is the SNOMED CT top hierarchy which shows the categories (see Fig. 8.7). Clinical findings is the categorization where clinical features including diagnoses are located. The schema is broad and covers both physical and conceptual entities. The format presented above and in these examples shows the name of the concept the distinction, the concept identifier, a local semantic type, and the number of children at the next level in the hierarchy.

This is the disease hierarchy with its base description logic (see Fig. 8.8). Note diseases have a semantic type of disorder which in SNOMED CT is called a distinction.

This is the hierarchy of diseases in SNOMED CT (see Figs. 8.9 and 8.10). You can see that the upper level of disorders has some road categories

- Disease (disorder) [64572001] [K] [60]
  - Hereditary disease (disorder) [32895009] [K] [16]
  - Disorder due to exposure to ionizing radiation (disorder) [85983004] [K] [62]
  - Vertiginous syndrome (disorder) [87118001] [K] [9]
  - Drug-related disorder (disorder) [87858002] [K] [250]
  - Communication disorder (disorder) [278919001] [K] [4]
  - Multisystem disorder (disorder) [281867008] [K] [73]
  - Developmental disorder (disorder) [5294002] [K] [15]
  - Environment related disease (disorder) [8504008] [K] [6]
  - Foreign body (disorder) (125670008) [K] [32]
  - Non-human disorder (disorder) [127326005] [K] [117]
  - Inflammatory disorder (disorder) [128139000] [K] [34]
  - Infectious disease (disorder) [40733004] [K] [48]
  - Disease of presumed infectious origin (disorder) [78885002] [K] [9]
  - Disorder characterized by pain (disorder) [373673007] [K] [48]
  - Chronic disease (disorder) [27624003] [K] [30]
  - Hematoma (disorder) [385494008] [K] [35]
  - Food poisoning (disorder) [75258004] [K] [5]
  - Nutritional disorder (disorder) [2492009] [K] [10]
  - Acute disease (disorder) [2704003] [K] [35]
  - Familial disease (disorder) [111941005] [K] [22]
  - Extraskeletal calcification (disorder) [237896000] [K] [50]
  - Neoplasm and/or hamartoma (disorder) [399981008] [K] [10]
  - Keratinizing cyst (disorder) [399999000] [K] [4]
  - Paraneoplastic syndrome (disorder) [49783001] [K] [11]
  - Disorder associated with menstruation AND/OR menopause (disorder) [106002000] [K] [11]
  - Maltreatment syndromes (disorder) [213015009] [K]
  - Angioedema and/or urticaria (disorder) [404177007] [K] [4]
  - Hyperviscosity syndrome (disorder) [11888009] [K]
  - Iatrogenic disease (disorder) [12456005] [K] [12]
  - Vomiting (disorder) [15387003] [K] [23]

Fig. 8.9 SNOMED CT disease hierarchy at the next level in the SNOMED CT hierarchy

and some leaf nodes such as "enterogenous cyst." This can either be a sign of imbalance in the terminology or more often that certain categories need more work (knowledge engineering) and therefore are less complete than other areas of the terminology.

The specific disorder *acute myocardial infarction* (AMI) (see Fig. 8.11) has a description logic definition that includes:

HasFindingSite *Myocardium structure*
HasCourse *Acute*
HasAssociatedMorphology *Acute infarct*

This definition should be unique in the terminology and serves to formally define the disorder *acute myocardial infarction*. The classifier can check to make sure that no other concept has this definition. It can also decide which other concepts have definitions indicative of the fact that

Fig. 8.10 SNOMED CT disease hierarchy at the next level in the SNOMED CT hierarchy continued from Fig. 8.9

- ⊞ Congenital disease (disorder) [66091009] [K] [133]
- ⊞ Substance abuse (disorder) [68214007] [K] [3]
- ⊞ Disease due to Arthropod (disorder) [66843000] [K] [48]
- ⊞ Disorder characterized by edema (disorder) [118654009] [K] [33]
- ⊞ Disorder by body site (disorder) [123946008] [K] [41]
- ⊞ Biphasic disease (disorder) [409701001] [K]
- ⊞ Disorder of cellular component of blood (disorder) [414022008] [K] [6]
- ⊞ Disorder of fetus or newborn (disorder) [414025005] [K] [13]
- ⊞ Disorder of hematopoietic cell proliferation (disorder) [414026006] [K] [7]
- ⊞ Disorder of immune function (disorder) [414029004] [K] [14]
- ⊞ Disorder of pigmentation (disorder) [414032001] [K] [35]
- ⊞ Obesity (disorder) [414916001] [K] [9]
- ⊞ Febrile disorder (disorder) [416113008] [K] [2]
- ⊞ Subacute disease (disorder) [19342008] [K] [12]
- ⊞ Mental disorder (disorder) [74732009] [K] [40]
- ⊞ Poisoning (disorder) [75478009] [K] [42]
- ⊞ Metabolic disease (disorder) [75934005] [K] [49]
- ⊞ Self-induced disease (disorder) [77434001] [K] [3]
- ⊞ Disorder of pregnancy (disorder) [173300003] [K] [16]
- · Enterogenous cyst (disorder) [204766008] [K]
- ⊞ Disorder of labor / delivery (disorder) [362972006] [K] [2]
- ⊞ Disorder of puerperium (disorder) [362973001] [K] [4]
- ⊞ Degenerative disorder (disorder) [362975008] [K] [91]
- ⊞ Sequela (disorder) [362977000] [K] [4]
- ⊞ Nutritional deficiency associated condition (disorder) [363246002] [K] [25]
- · Obesity associated disorder (disorder) [363247006] [K]
- ⊞ Traumatic AND/OR non-traumatic injury (disorder) [417163006] [K] [37]
- ⊞ Propensity to adverse reactions (disorder) [420134006] [K] [3]
- ⊞ AIDS-associated disorder (disorder) [420721002] [K] [58]
- ⊞ Hypersensitivity disorder (disorder) [421976005] [K] [4]

Fig. 8.11 The description logic definition for *acute myocardial infarction* in SNOMED CT

- ⊟ Acute myocardial infarction (disorder)
  - ⊞ Synonyms
  - ⊞ More specific concepts
  - ⊞ Broader Concepts
  - ⊟ Finding site
    - · Myocardium structure (body structure) [74281007] [M]
  - ⊞ Episodicity
  - ⊟ Course
    - · Acute (qualifier value) [53737009] [M]
  - ⊟ Associated morphology
    - · Acute infarct (morphologic abnormality) [55470003] [M]
  - ⊞ Severity

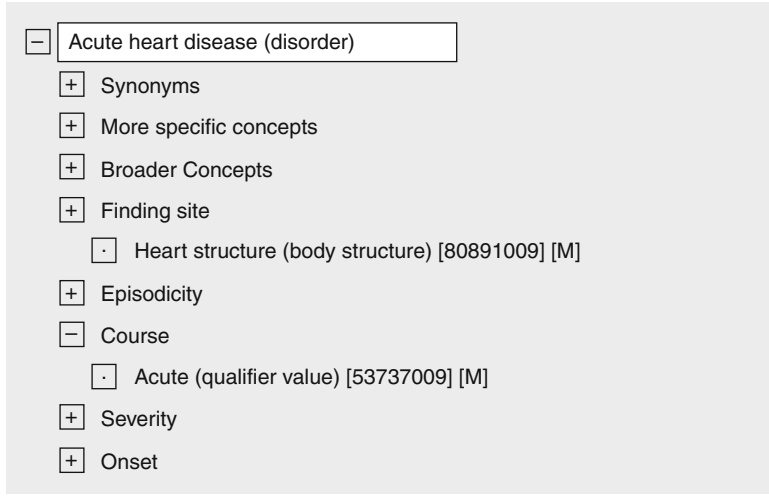


Fig. 8.12 The description logic definition for *acute heart disease* in SNOMED CT

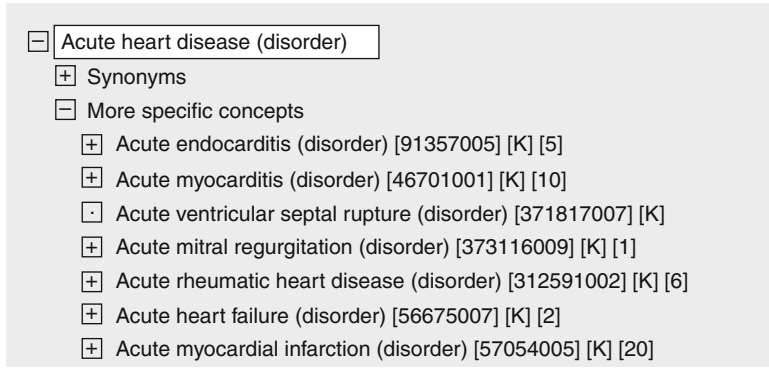


Fig. 8.13 More specific concepts related to *acute heart disease* in SNOMED CT

AMI is a child of that concept. In this case that would include two concepts: *myocardial infarction* and *acute heart disease*. Acute heart disease has the description logic definition shown in Fig. 8.12:

As myocardium structure Isa heart structure and acute course Isa acute course (this property is called reflexivity), AMI is a type of acute heart disease.

As you can see the broader concepts and the more specific concepts, assignments are reciprocal and consistent (see Fig. 8.13). If A is broader than B, then B is more specific than A.

All in all, SNOMED CT is the largest and most expressive of the healthcare terminologies. It is a compositional system and has a description logic underpinning its representation. In one study, it was shown that SNOMED CT could represent 92.3% of clinical problems when used as a compositional system [7]. The same article noted that the coverage dropped to 51.4% if SNOMED CT was used only as a precoordinated terminology. When we compared these results, with and without composition, the compositional nature of SNOMED CT significantly contributed to its total coverage of clinical problems ($p < 0.001$). Another study of mapping two precoordinated terminologies demonstrated the value of SNOMED CT as a reference terminology [8]. In yet another study, each precoordinated terminology was mapped to SNOMED CT, and in addition to a more straightforward lexical matching, the two precoordinated terminologies were mapped using their more granular SNOMED CT associations leading to an increase in coverage from 33% to 95% [9].

NDF-RT

Part of the back end of the CPRS drug ordering system is the VHA National Drug File (NDF), a nationally maintained drug resource that local VHA pharmacies must map to in order to enable the existing drug–drug interaction detection application and to support Centralized Mail-Out Pharmacy (CMOP) prescriptions. However, individual medical center pharmacies maintain their own site-specific drug files and formularies that may not be computationally comparable with others. When a patient is seen at multiple VA medical centers, drug system incompatibilities could result in medication errors and missed opportunities for data aggregation.

NDF-RT (National Drug File – Reference Terminology) is an enhancement of NDF. First, because it will be deployed as a reference terminology, it can be used by all VA medication order entry and medication display sites in a standard way. Second, it is designed to support aggregations of "similar" drugs, for example, drugs containing the same active ingredient, so that decision support and analytical applications can be created, maintained, and used productively. Third, because it is being developed using federal resources, it will be free of the intellectual property restrictions associated with some proprietary drug information systems; if enforced, these restrictions can impede evolutionary deployment, reuse, and interoperation with other agencies and enterprises.

The NDF-RT is a terminology which is based upon a description logic representation. The content and semantics are designed to facilitate the representation of drugs at the level of the clinical drug. This level of specificity is to be useful for prescribing and ordering drugs (where in some circumstances, such as a consultation where a drug may be prescribed, but not ordered, they differ).

Dr. Steven Brown was the principal investigator on the NDF-RT project. He and Dr. Michael Lincoln were major contributors to the richness of this medication terminology. In fiscal year 2001, the Department of Veterans Affairs (VA) Veterans Health Administration (VHA) provided healthcare to 4.1 million veterans and dependents in the form of 43 million outpatient visits, 573,000 inpatient admissions, and 167 million prescriptions (as 30-day equivalents). VHA has developed and deployed a variety of electronic tools to assist clinicians, including VISTA (Veterans Integrated Service and Technology Architecture) [10, 11], CPRS (Computerized Patient Record System) [12, 13], BCMA (Bar Code Medication Administration), [14, 15] and others.

VHA is continually looking for ways to use information technology (and other tools) to improve care quality, promote patient safety, and reduce costs. Reference terminologies and terminology services that permit retrospective and real-time aggregation and sophisticated decision support are one such area under investigation. Formal terminologies are also being evaluated as a way to reduce maintenance and mapping effort [5, 16]. VHA's initial reference terminology project is NDF-RT [17, 18], a formalization of the National Drug File. Other reference terminologies will be deployed under the VHA Enterprise Reference Terminology project.

NDF-RT uses a description-logic-based reference model which includes a defined set of abstractions denoting levels of description for drug products (based on work performed within the Health Level 7 (HL7) Vocabulary Technical Committee) [19], a set of hierarchical and definitional relationships, and sets of nondefinitional properties used at each hierarchical level to capture associated details. The model includes hierarchies for chemical structure, mechanism of action, physiologic effect [18], and therapeutic intent [17]. As of September 2003, NDF-RT was the final phases of expert review by doctors of clinical pharmacology. The most recent version of NDF-RT includes 4202 active ingredients (including salt forms) and 108,112 National Drug Code (NDC) level products. Role definition counts (including inferred roles) are 118,504 mechanism_of_action roles, 119,095 physiologic_effect roles, 123, 379 may_treat roles, 52,827 may_prevent roles, and 5522 may_diagnose roles.

A number of papers detailing desirable characteristics of terminologies have been published in the past 5 years. For example, in 1998, Cimino

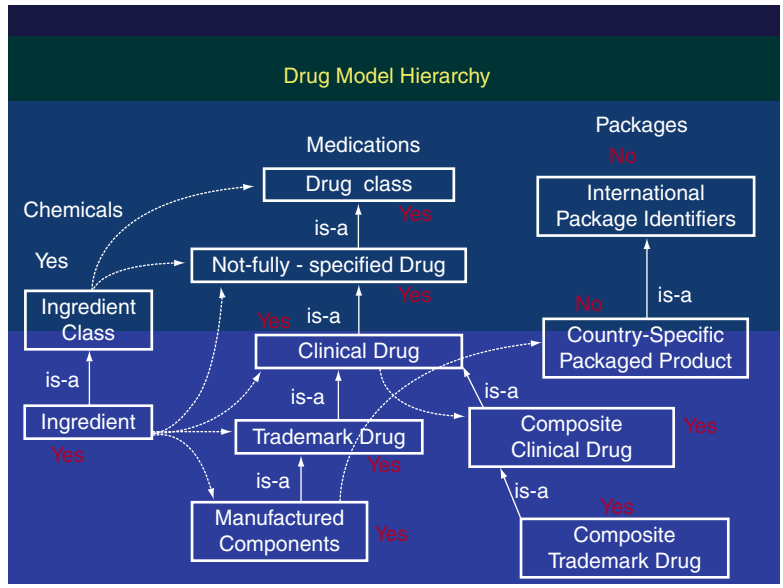


Fig. 8.14 Drug terminology reference information model

[20] described 12 "desiderata" synthesized from the literature of medical vocabulary research. Two additional publications [21, 22] by Elkin et al. advance our understanding of terminology quality indicators even further. In the standard world, ASTM E 2087-00 [23] and ISO TS17117 [24] make significant contributions toward defining the quality criteria for terminologies. Each of these works acknowledges the importance of content coverage. One of the key principles of terminology is (to paraphrase the old adage) "content, content, and more content." [20].

Further we will look at the component model of drugs (see Fig. 8.14) to determine if the different parts of a clinical drug can be individually identified for all clinical drugs (Form, Dose with Units, Route, Frequency, Duration, # Disp, Refills). International package identifiers need to be added to the NDF-RT to satisfy the original model.

NDF-RT defines a clinical drug and its relationship to a packaged drug along with its indications, clinical effects, physiological effects, and mechanism of action. The generic form of these relations can be seen in Fig. 8.15, and a specific set of populated elements as seen in Fig. 8.16 show how these relate in the context of a real clinical drug the angiotensin converting enzyme (ACE) inhibitor captopril.

This is a well-formed terminology. Of note both the physiological effects and mechanism of actions hierarchies have been adopted as US national standards by NCHS.

RxNorm [25]

RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software, including those of First Databank, Micromedex, MediSpan, Gold Standard Alchemy, and Multum. By providing links between these vocabularies, RxNorm can mediate messages between systems not using the same software and vocabulary. RxNorm is created by the National Library of Medicine (NLM) of the National Institutes of Health (NIH). Dr. Stuart Nelson is the creator of RxNorm which was described in collaboration with Dr. Steven Brown of the Veterans Administration [28].

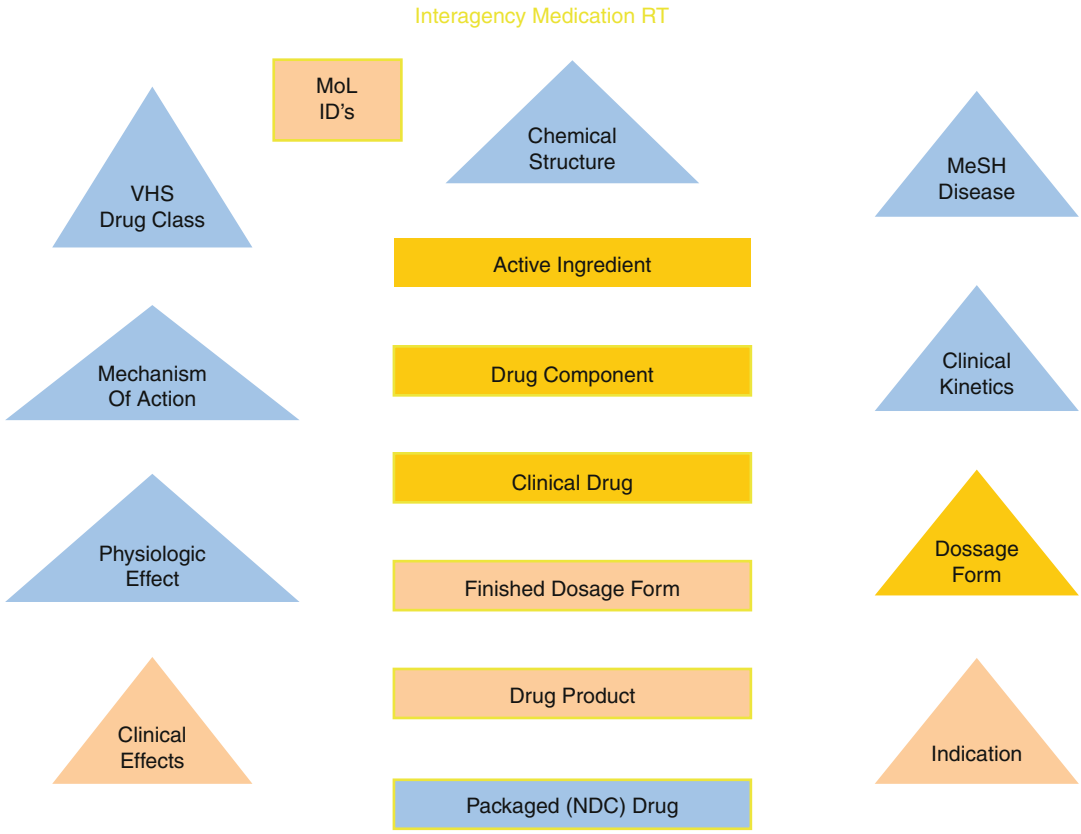


Fig. 8.15 NDF-RT reference data model – Clinical Effects were not modeled, and there are some clinical kinetics, but they have only been partially modeled to date. Otherwise, NDF-RT contains the type of information identified on this slide

In late 2001, the NLM began experimenting with the representation of clinical drugs within the Unified Medical Language System. There were several reasons for wanting to represent drugs: first, there was the suspicion that within the Metathesaurus, there was considerable synonymy missed as clinical drugs were named; second, traditional methodologies of identifying missed synonymy in the UMLS did not seem to be working for clinical drugs; third, there was hope that developing a new set of models in the domain of clinical drugs might lead to improved interoperability of drug terminologies; fourth, the area of clinical drugs was seen as important and relevant with respect to patient safety; and fifth, there was a growing consensus in the HL7 vocabulary technical committee, based on their work on the adoption of NDF-RT's clinical drug model, as to what should be the appropriate model for clinical drugs. The HL7 model was based on what a clinician would order, in a form that would be appropriately sent to the pharmacy. The dose form should represent how the drug would be administered to a patient, as opposed to the form in which the manufacturer had supplied the drug. The clinical drug form was seen as clearly distinct from the choices the pharmacy might make in fulfilling that order.

The RxNorm project had as stepwise approach to clinical drug representation. Step one was to define a Semantic Normal Form (SNF) for the representation of clinical drugs. SNFs for clinical drugs are canonical representations that are defined by their active ingredients, strengths, and orderable dose forms. The SNF instantiates relationships between concepts in attribute–value

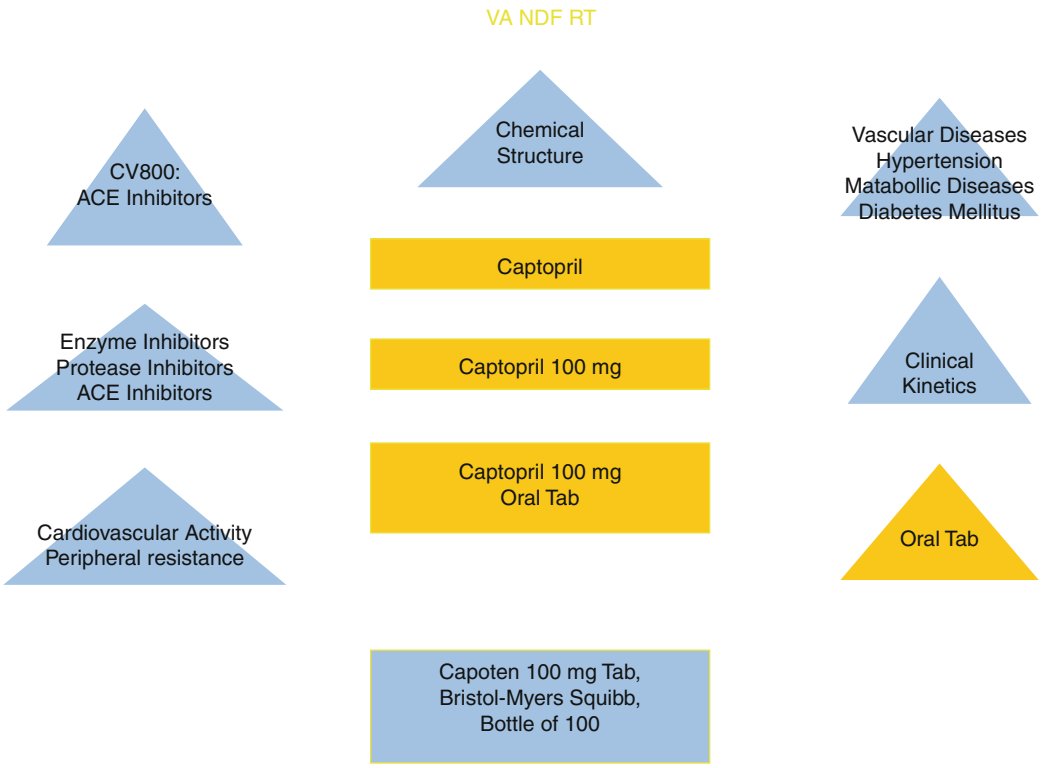


Fig. 8.16 Example of the populated data model for NDF-RT

pairs. SNFs for clinical drugs use a set of standardized (generic) ingredient names, units, and dose forms, and a set of rules for expressing strength in a set of standard units.

There are two SNFs created as UMLS concepts for every clinical drug. The SNF Drug Component (SCDC) has the form:

CUI|ShortName|ActiveIngredient|PreciseIngredient|Basis|Strength|Units|Notes

For example, the following are SCDCs:

C0111111|APAP|Acetaminophen|Acetaminophen|B|325|MG|Component example#1

C0123456|Codeine|Codeine Phosphate|Codeine|P|30|MG|Component example#2

The SNF Clinical Formulation (SCD) has the form:

CUI|MetaID|ShortName|Component1/Component2/…|OrderableDoseForm|Notes

For example, the two SCDCs above can be combined to form the following:

C0654321|ACETAMINOPHEN 325 MG/CODEINE 30 MG Oral Capsule|C0111111/C0123456|Oral Tablet|CF example

How RxNorm Is Structured?

An RxNorm clinical drug name reflects the active ingredients, strengths, and dose form comprising that drug. When any of these elements vary, a new RxNorm drug name is created as a separate concept (explained below). Thus, an RxNorm name should exist for every strength and dose of every available combination of clinically significant ingredients.

Connections, in the form of predefined relationships, exist among the components of RxNorm and between RxNorm concepts and concepts derived from other vocabularies contained in the UMLS Metathesaurus.

RxNorm data are distributed in either the Metathesaurus Relational (MR) or Rich Release Format (RRF) tables. The tables that may be of particular relevance are:

- RXNCONSO, Concept and Source Information
- RXNREL, Relationships
- RXNSAT, Attributes
- RXNSTY, Semantic Type

Details of UMLS and RxNorm Structure

- Complete information on the structure of the UMLS system, its data elements, and its tables can be found in the UMLS Reference Manual, available online at http://www.ncbi.nlm.nih.gov/books/NBK9676/
- In part because of their different schedules of updates, RxNorm file structure differs in some ways from that of the UMLS. Details on the relation between RxNorm and UMLS files may be found at http://www.nlm.nih.gov/research/umls/rxnorm/docs/index.html.
- The RxNorm Navigator (RxNav) is an interactive graph built on the model of the figure in the section on RxNorm Relationships below. It allows you to query the RxNorm database by any of its components. RxNav can be found at http://mor.nlm.nih.gov/download/rxnav

The Elements of a Normalized Form of a Clinical Drug

RxNorm follows a standard format in the naming of clinical drugs. Drugs identified in other vocabularies are linked to a normalized name prepared according to RxNorm's naming conventions.

The normalized form of the clinical drug name may be thought of as being composed of a number of components; each component is a concept. Each element of the normalized form can be identified by its value in the TTY [Term Type] field of the RXNCONSO file. The possible range of values are as follows (Table 8.10):

A term type not listed here is OCD (i.e., obsolete clinical drug).

One final element found in some of the normal forms is the quantity factor. The quantity factor is not represented as a separate term type.

A Concept Orientation: RxNorm's Links to Other Vocabularies

Like the UMLS Metathesaurus as a whole, RxNorm is organized by concept. A concept is a collection of names identical in meaning at a specified level of abstraction. It serves as a means whereby strings of characters from disparate sources may be taken to name things that are the same.

For example, "Accuneb, 0.042% inhalation solution and Albuterol 0.417 MG/ML Inhalant Solution [Accuneb]" name the same concept. Where a normalized form exists in RxNorm, it is designated as the preferred form of the drug name (this is denoted by its association with the TS [Term Status] field in RXNCONSO). The concept is assigned an RxNorm concept unique identifier (RXCUI) which in this case is 575803. This RXCUI always denotes the same concept, regardless of the form of the name and regardless of what table it resides. Drugs whose names map to the same RXCUI are the same drug and should be identical in their ingredients, strengths, and dose forms. Conversely, drugs that differ in any of these parameters are to RxNorm conceptually distinct and will have different RXCUIs.

"Acetaminophen 500 MG Oral Tablet" and "Acetaminophen 500 MG Oral Tablet [Tylenol]" name two different concepts based on the name Tylenol being included with the later term, with RXCUIs 198440 and 209459, respectively. The first of these concepts, 198440, has the relationship "has_tradename" to the second concept, 209459, and the second concept has the reciprocal relationship "tradename_of" to the first concept.

Two brand name drugs for the same generic components would refer to different concepts. For example,

"Fluoxetine 20 MG Oral Capsule [Prozac]" and

"Fluoxetine 20 MG Oral Capsule [Sarafem]"

Table 8.10 Drug fields and examples of their content

TTY	Name	Definition	Example(s)
IN	Ingredient	A compound or moiety that gives the drug its distinctive clinical properties. The preferred name is usually the USAN name	Fluoxetine, insulin, isophane, human gentamicin sulfate (USP)
PIN	Precise ingredient	A specified form of the ingredient that may or may not be clinically active. Most precise ingredients are salt or isomer forms	Fluoxetine hydrochloride
MIN	Multiple ingredients	Two or more ingredients created from SCDF. In rare cases when IN/PIN or PIN/PIN combinations of the same base ingredient exist, created from SCD	Fluoxetine/olanzapine
DF	Dose form	A complete list of dose forms can be found in Appendix 2 of the RxNorm documentation	Topical solution, oral tablet
SCDC	Semantic clinical drug component	Ingredient plus strength – see section on Rules and Conventions, below, for units of measurement and for rules pertaining to the calculation of strengths	Fluoxetine 4 MG/ML
SCDF	Semantic clinical drug form	Ingredient plus dose form	Fluoxetine oral solution
SCD	Semantic clinical drug	Ingredient plus strength and dose form	Fluoxetine 4 MG/ML oral solution
BN	Brand name	A proprietary name for a family of products containing a specific active ingredient	Prozac
SBDC	Semantic branded drug component	Branded ingredient plus strength	Fluoxetine 4 MG/ML [Prozac]
SBDF	Semantic branded drug form	Branded ingredient plus dose form	Fluoxetine oral solution [Prozac]
SBD	Semantic branded drug	Ingredient, strength, and dose form plus brand name	Fluoxetine 4 MG/ML oral solution [Prozac]
SY	Synonym of another TTY	Given for clarity	Prozac 4 MG/ML oral solution
TMSY	Tall Man lettering synonym of another TTY	Given to distinguish between commonly confused drugs	Fluoxetine 10 MG oral capsule [Prozac]
BPCK	Brand name pack	Branded drug delivery device	{12 (ethinyl estradiol 0.035 MG/norethindrone 0.5 MG oral tablet)/9 (ethinyl estradiol 0.035 MG/norethindrone 1 MG oral tablet)/7 (inert ingredients 1 MG oral tablet)} pack [Leena 28 Day]
GPCK	Generic pack	Generic drug delivery device	{11 (varenicline 0.5 MG oral tablet)/42 (varenicline 1 MG oral tablet)} pack

Linking from RxNorm to DailyMed

The HL7 standard Structured Product Labels from the US FDA that describe each approved clinical drug in the United States are available on the NLM DailyMed website and may be accessed using RxNorm by constructing a DailyMed URL from the SPL_SET_ID attribute associated to the RXCUI of an RxNorm concept (Table 8.11). The structure of the URL is as follows:

http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=<insert SPL_SET_ID here>

The suffix to this URL is the value of the SPL_SET_ID attribute found in the RXNSAT table. For example, for RXCUI 102166, this query, denoted below, provides the following results:

select c.rxcui, c.str, s.atv from rxnconso c, rxnsat s where c.rxcui = s.rxcui and s.atn = 'SPL_SET_ID' and c.rxcui = '102166 and c.sab = 'RXNORM' and c.suppress = 'N'

Thus, using the results for Robinul, the URL to the DailyMed label would be: http://dailymed.nlm.nih.gov/dailymed/lookup.cfm?setid=bd65ee5e-2000-423c-b0a6-72eb213455c4

Table 8.11 The RXCUI identifier and examples of its associated content

RXCUI	STR	ATV
102166	Glycopyrrolate 2 MG Oral Tablet [Robinul]	bd65ee5e-2000-423c-b0a6-72eb213455c4
102166	Robinul 2 MG Oral Tablet	bd65ee5e-2000-423c-b0a6-72eb213455c4

RxNorm Relationships

Relationships between concepts in RxNorm are reciprocal. Each direction of the relationship is represented as a separate row in RXNREL. A clinical drug consists of a set of precoordinated components, and the precoordination in turn constitutes the clinical drug.

RxNorm Concepts and Relationships Diagrams

RXCUI is the concept unique identifier. TTY is the term type.

Relationships at the level of Clinical Drug showing the components of each drug and its relationships to its ingredients are shown in Fig. 8.17.

Relationships at the level of drug packaging to clinical drug are shown in Fig. 8.18.

RxNorm contains the following relationships:

- constitutes/consists_of
- contains/contained_in
- dose_form_of/has_dose_form
- form_of/has_form
- includes/included_in
- ingredient_of/has_ingredient
- isa/inverse_isa
- precise_ingredient_of/has_precise_ingredient
- tradename_of/has_tradename

Rules and Conventions Used to Generate RxNorm Data

Naming Conventions

The SCD – the semantic clinical drug – also called the normalized form of the generic drug name contains the ingredient(s), the strength, and the dose form, represented in that order. The components and forms of an SCD, its SCDCs and SCDFs, contain the ingredient and strength and the ingredient and dose form, respectively. The SBD follows a similar naming convention; it also contains the brand name in brackets at the end of the name.

The ingredients named in the SCD, SBD, etc., are all active ingredients. Thus, in the example clinical drug shown in the above figure, cetirizine is used as the ingredient name. Though "cetirizine" and "cetirizine dihydrochloride" are separate concepts, the normalized form of the drug name does not include the precise ingredient name since in this case, the difference is without clinical significance. This can be confusing, and it is important to note since this terminology is intended to serve up the clinical drug name for clinical use, most clinicians would only want to

Fig. 8.17 Relationships at the level of clinical drug showing the components of each drug and its relationships to its ingredients

use the name cetirizine when prescribing this medication.

RxNorm makes no distinction between amoxicillin trihydrate, amoxicillin monosodium salt, and amoxicillin potassium salt, because the differences among them are not commonly clinically significant. When there are significant differences among preparations, as is the case with "Penicillin G, Benzathine" vs. "Penicillin G, Procaine," the entire compound name (the PIN) is always included as the ingredient. Here the drugs have considerably different half lives with benzathine penicillin working for 2 weeks in the patient while procaine penicillin only working for 1 week in duration.

Brand Name Drugs

Distinct concepts are created in RxNorm for brands whose formulations (i.e., whose aggregates of ingredients) are distinct.

This can lead to some unexpected results, for example, Bactrim and Bactrim DS both contain sulfamethoxazole and trimethoprim (and in the same proportions relative to each other); the DS indicates that its product is twice as strong as the other. RxNorm Records for both products link to the same BN code.

In the case of Claritin (loratadine) vs. Claritin D (loratadine with pseudoephedrine), the "D" indicates that there is an additional ingredient. RxNorm, in this case, contains distinct BNs for Claritin and for Claritin D.

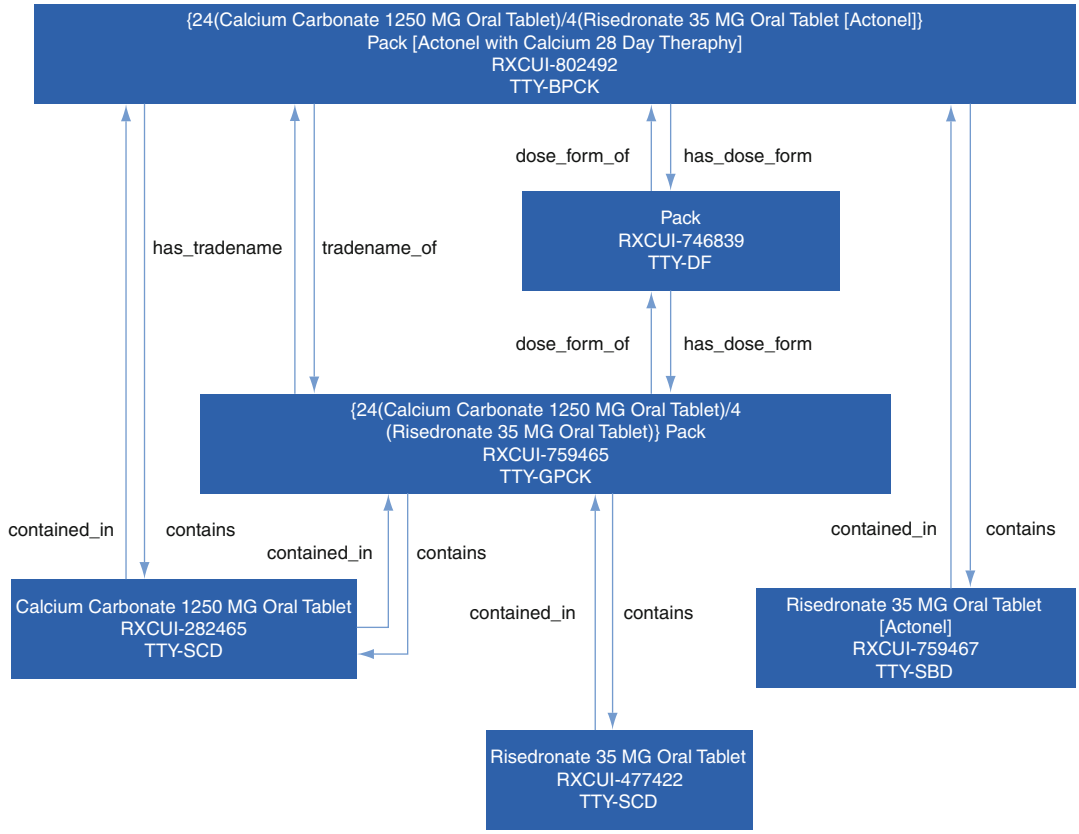


Fig. 8.18 Relationships at the level of drug packaging to clinical drug

If a drug contains more than one component and the dosing proportions are different, this will cause RxNorm to generate a separate BN. For example, Advair contains fluticasone and salmeterol. The amount of salmeterol remains constant; however, the dose of fluticasone varies in the different clinical drugs. RxNorm creates several BNs for Advair, including, but not limited to, Advair Diskus 250/50 and Advair Diskus 50/50.

Strengths

The strengths are assigned to the active ingredient. In drugs where there is more than one active ingredient, there will be a separate strength value associated with each ingredient, as in the example SCD below:

> "Ascorbic Acid 100 MG/Calcium Carbonate 625 MG/Ferrous Fumarate 122 MG/Folic Acid 1 MG Oral Tablet"

Here, the SCD has the relationship "consists_of" assigned to each of the ingredient-strength pairs (i.e., SCDCs) that are separated by slash marks.

Strengths are always expressed to three significant digits. Some nearly equal strengths that may be expressed differently in other drug vocabularies are treated by RxNorm as equivalent. In the case of most drugs, the active ingredient will be one of the ingredients (IN). Some drugs that contain a mixture of salts that each have a significant and different clinical action will require a separate IN. For example, Adderall contains two ingredients: amphetamine and dextroamphetamine. There are six forms of these two ingredients within the clinical drug, four of which are clinically active. The RxNorm SCD creates separate names for each of these salts with its individual strength.

"Amphetamine Aspartate 1.25 MG/Amphetamine Sulfate 1.25 MG/Dextroamphetamine Saccharate 1.25 MG/Dextroamphetamine Sulfate 1.25 MG Extended Release Capsule"

For small inorganic molecules, the strength will be expressed in terms of the salt given as the different salts have different absorptions and therefore clinical uses. For example, an oral tablet that contains 1250 mg of calcium carbonate and 500 mg of free calcium will appear as:

"Calcium Carbonate 1250 MG Oral Tablet"

Units of Measurement

In RxNorm, units are used to standardize the expressions of strength. Strengths that are expressed as ratios assume a value of 1 for the denominator. Therefore, 100 mg in 5 ml would be expressed as 20 mg/ml.

RxNorm uses the following units of measurement:

- CELLS
  - Cells
- MEQ
  - Milliequivalent
- MG
  - Milligram
- ML
  - Milliliter
- UNT
  - Unit
- %
  - Used only in association with gases, other percentages are converted into ratios.
- ACTUAT
  - Actuation. Refers to a measured dose per activation of a dispensing devise; e.g., in an inhaler, the strength of the clinical drug is given by how much is dispensed with each actuation. This unit appears only in ratios.

The following ratios of units are used in RxNorm:

- CELLS/ML
- MEQ/MG
- MEQ/ML
- MG/ACTUAT
- MG/MG
- MG/ML
- ML/ML
- PNU/ML
- UNT/MG
- UNT/ML

Conversion of Units

The rules followed for RxNorm standard units are:

- Standard conversion factors are used between metric units.
- One liquid ounce is equivalent to 30 ml or 240 ml per 8 ounces.
- A grain is equal to 65 milligrams.

RxNorm calculates the concentration that determines the strength based upon the minimal diluents that can be used. For example, RxNorm would use 3 ml in the case of a drug that can be dissolved in 3 to 5 ml of diluent. For drugs that have multiple dilution steps, only the initial dilution is used to determine the strength. For example, a vial containing 50 mg of a drug to be dissolved in 2 ml of water which then is added to an IV solution is expressed as having strength of 25 mg/ml.

Reformulated Drugs

Reformulated drugs are drugs whose ingredients have been changed by the manufacturing company but continue to be identified by the original brand name. A new BN is created in RxNorm to designate the reformulation and provide the date of the change. Sample SBDs are shown below:

"Dihydroxyaluminum Sodium Carbonate 334 MG [Rolaids]" became

"Calcium Carbonate 550 MG/Magnesium Hydroxide 110 MG [Rolaids Reformulated Aug 2006]"

Synonym Use

Because the names and strengths of each component are listed, normal forms may sometimes grow to inordinate lengths. This will be true of multivitamins or ionic solutions such as Lactated Ringer's Irrigation Solution.

In such cases, synonyms (TTY = "SY") will be created in RxNorm as more manageable forms of the name. SY atoms can be created for normal forms of the type GPCK, BPCK, SCD, and SBD. Each of these normal forms can have multiple SY atoms in their concepts.

Strength Expressed as Precise Ingredient

If RxNorm receives a string from one of its source vocabularies with the strength expressed in terms of the precise ingredient, this will be noted as an attribute in RXNSAT. See example below:

Amiodarone hydrochloride 200 MG Oral Tablet

Codes and CUIs

The code field values (e.g., in RXNCONSO) are obtained from the source vocabularies.

Table 8.12 indicates the fields from the source vocabularies from which the RxNorm Code is taken:

For the two sources listed in Table 8.13, the field from which the code is derived is determined by the term type.

The last two MMSL term types are loaded into RxNorm as "NoCode".

NLM is now associating US Food and Drug Administration (FDA) generated unique ingredient identifiers (UNIIs) with the RxNorm (SAB = RXNORM) term type IN atoms. This association is made utilizing an exact case-insensitive string match to the RxNorm ingredient string from the official FDA substance list. These UNII codes can be found in the RXNSAT.RRF file and are denoted as values of the attribute ATN = "UNII_CODE". The UNII is a nonproprietary, free, unique, unambiguous, nonsemantic, alphanumeric identifier based on a substance's molecular structure and/or descriptive information.

Cardinality

If a BN has more than one IN, this is denoted as an attribute of the BN with the value "multi" and can be found in the table RXNSAT.

Updates

The full set of files is included in the UMLS Metathesaurus. The UMLS Metathesaurus is updated two to three times a year, while RxNorm is available as a full update on a monthly basis. An update containing data from the DailyMed website along with any newly approved drugs is available in weekly releases. For interested users, the RxNorm update files will be made available through the UMLS Knowledge Source Server.

Obsolete Records

Obsolete records can be identified in three ways:

1. When one of RxNorm's source vocabularies drops a clinical drug name. That is when a drug name had been used in a previous version of that source vocabulary, however, it is not found in the most recent version. Then the old clinical drug records are instantiated with the term type OCD (for obsolete clinical drug) in RXNSTY. To show that the code is now not from the source vocabulary, the record is updated with RxNorm as the source (SAB field), but retains the original SAB, VSAB, TTY, and code as attributes in RXNSAT. All existing relationships to RxNorm records are persisted.
2. When a clinical drug disappears from the US market, the RxNorm SCD should correspond only to a term type of OCD. Once this happens, the drug name is flagged with an "O" (meaning obsolete) in the suppress field in RXNCONSO.
3. If during the process of resynchronization with the UMLS it is found that there is more than one RxNorm record for the same concept, then one of the records is marked as the preferred drug record, and the others are archived. The archive file is held in a file named RXNATOMARCHIVE.

Table 8.12 Information used in RxNorm stratified by the source of that information

Source	Source field from which code drawn
FDA NDC Directory (MTHFDA)	Listing_Seq_No
First DataBank (NDDF)	GCN_Seq_No
Micromedex RED BOOK (MMX)	GFC_Code
SNOMED CT	SNOMED Concept ID
RxNorm	RXCUI
VHA National Drug File (VANDF)	VUID

Table 8.13 The term type can be used to determine where the information is stored within RxNorm

Source	Term type	Field from which code drawn
FDA Structured Product Labeling (MTHSPL)	DP	Product_ID
FDA Structured Product Labeling (MTHSPL)	SB	UNII
Multum MediSource Lexicon (MMSL)	BD	Main_Multum_Drug_Code
Multum MediSource Lexicon (MMSL)	BN	Brand_Code
Multum MediSource Lexicon (MMSL)	CD	Main_Multum_Drug_Code
Multum MediSource Lexicon (MMSL)	GN	Drug_ID
Multum MediSource Lexicon (MMSL)	IN	Active_Ingredient_Code
Multum MediSource Lexicon (MMSL)	MS	"NOCODE"
Multum MediSource Lexicon (MMSL)	SC	"NOCODE"

Downloading RxNorm

RxNorm files can be obtained using the NLM download server:

http://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html

Nursing Terminologies

The nursing informatics community has been very engaged and thoughtful in their terminological efforts. In many ways they have led the way for the future of medicine. Specifically, nursing has created representational schemes for goal statements which measure health outcomes and present targets for good care of patients. There are many worthy terminological efforts in nursing. The American Nursing Association has recognized the following terminologies as of August 2010:

1. **NANDA – Nursing Diagnoses, Definitions, and Classification** 1992
   2009–2010 Version now available
   NANDA International
   P.O. Box 157
   Kaukauna, WI 54130–0157
   Phone: 1-920-344-8670
   Email: nanda@nanda.org
   Website: www.nanda.org
2. **Nursing Interventions Classification System (NIC)** 1992
   Sue Moorehead, PhD, RN, Center Director
   The Center for Nursing Classification and Clinical Effectiveness
   University of Iowa
   College of Nursing, 458 Nursing Building
   Iowa City, IA 52242–1121
   Phone: 319-335-7051
   Fax: 319-335-6820
   Website: www.nursing.uiowa.edu/excellence/nursing_knowledge/clinical_effectiveness/index.htm
   (NIC/NOC can be obtained from the same source)
3. **Clinical Care Classification (CCC)** 1992
   Formerly Home Health Care Classification (HHCC)
   Virginia K. Saba, EdD, RN, FAAN, FACMI
   Georgetown University School of Nursing
   3700 Reservoir Road, NW
   Washington, DC 20007
   Phone: 703-521-6132 (h)
   Fax: 202-687-5553
   Website: www.sabacare.com
4. **Omaha System** 1992
   Karen S. Martin, MSN, RN, FAAN
   Martin Associates
   2115 S. 130th Street
   Omaha, NE 68144
   Phone: 402-333-1962
   Email: martinks@tconl.com
   Website: www.omahasystem.org
5. **Nursing Outcomes Classification (NOC)** 1997
   Sue Moorehead, PhD, RN, Center Director
   Center for Nursing Classification and Clinical Effectiveness
   University of Iowa
   College of Nursing, 458 Nursing Building
   Iowa City, IA 52242–1121
   Phone: 319-335-7051
   Fax: 319-335-6820

Website:
www.nursing.uiowa.edu/excellence/nursing_knowledge/clinical_effectiveness/index.htm
(NIC/NOC can be obtained from the same source)

6. **Nursing Management Minimum Data Set (NMMDS)** 1998
Connie Delaney, PhD, RN, FAAN, FACMI
Co-PI, NMMDS
School of Nursing
5–160 Weaver Densford Hall
The University of Minnesota
308 Harvard Street SE
Minneapolis, MN 55455
Phone: 612-624-5959
Fax: 612-624-3174
Email: Delaney@umn.edu
Website: http://www.nursing.umn.edu/ICNP/USANMMDS/home.html
7. **Perioperative Nursing Data Set (PNDS)** 1999
Carole Peterson, Lead
Sharon Giarriczo-Wilson
Association of PeriOperative Registered Nurses
2170 South Parker Road, Suite 300
Denver, CO 80231–5711
Phone: 800-755-2676 ext. 392
Phone: 800-755-2676 ext. 472
Email: pnds@aorn.org
Website: www.aorn.org or www.aorn.org/PracticeResources/PNDSAndStandardizedPerioperativeRecord/
8. **SNOMED CT** 1999
Cynthia B. Lundberg, BSN, RN
Clinical Informatics Educator
SNOMED Terminology Solutions
A Division of the College of American Pathologists
500 Lake Cook Road, Suite 355
Deerfield, IL 60015
Phone: 1-800-323-4040, ext. 7673 or 847-832-7673
Fax: 847-832-8335
Email: clundbe@cap.org
Website: www.ihtsdo.org/snomed-ct/
9. **Nursing Minimum Data Set (NMDS)** 1999
Connie Delaney, PhD, RN, FAAN
School of Nursing
5–160 Weaver Densford Hall
The University of Minnesota
308 Harvard Street SE
Minneapolis, MN 55455
Phone: 612-624-5959
Fax: 612-624-3174
Email: Delaney@umn.edu
Website: http://www.nursing.umn.edu/ICNP/USANMDS/home.html
10. **International Classification for Nursing Practice (ICNP®)** 2000
Amy Coenen, PhD, RN, FAAN, Associate Professor
Director, International Classification for Nursing Practice (ICNP®)
International Council of Nurses
University of Wisconsin – Milwaukee
College of Nursing
P.O. Box 413
Milwaukee, WI 53201–0413
Phone: 414 229–5146
Fax: 414 229–6474
Email: coenena@uwm.edu
Website: http://www.icn.ch/icnp.htm
Amy Amherdt, ICNP®
Administrative Assistant
Phone: 414-229-5501
Fax: 414-229-6474
Email: aamherdt@uwm.edu
11. **ABC Codes** 2000
Melinni Giannini, President
ABC Coding Solutions
P.O. Box 20069
Albuquerque, NM 87154
Tel: 505-875-0001
Toll Free: (877) 621–5465
Fax: 505-875-0002
Email: Melinna.Giannini@alternativelink.com
Website: www.abccodes.com
12. **Logical Observation Identifiers Names and Codes (LOINC®)** 2002
Susan Matney, RN, MS
Chair, Nursing Clinical LOINC Subcommittee
Susan Matney, MSN, RN
Department of Biomedical Informatics, University of Utah

26 South 2000 East
Salt Lake City, Utah 84112
Email: susan.matney@utah.edu
Cell Phone: 801-585-9871
Fax: 801-581-4297
Website: http://loinc.org

As we do not have space to discuss all of these terminologies, we will use ICNP as an example of a well-formed nursing terminology.

ICNP [26]

Vision

ICNP® is an integral part of the global information infrastructure informing healthcare practice and policy to improve patient care worldwide.

Strategic Goals

- Serve as a major force to articulate nursing's contribution to health and healthcare globally
- Promote harmonization with other widely used classifications and the work of standardization groups in health and nursing

Benefits

- Establishes an international standard to facilitate description and comparison of nursing practice
- Serves as a unifying nursing language system for international nursing based on state-of-the-art terminology standards
- Represents nursing concepts used in local, regional, national, and international practice, across specialties, languages, and cultures
- Generates information about nursing practice that will influence decision making, education, and policy in the areas of patient needs, nursing interventions, health outcomes, and resource utilization
- Facilitates the development of nursing data sets used in research to direct policy by describing and comparing nursing care of individuals, families, and communities worldwide
- Improves communication within the discipline of nursing and across other disciplines
- Encourages nurses to reflect on their own practice and influence improvements in quality of care

The ICNP® is a unified nursing language system. It is a compositional terminology for nursing practice that facilitates the development of and the cross-mapping among local terms and existing terminologies.

ICNP® Elements

- Nursing phenomena (nursing diagnoses)
- Nursing actions
- Nursing outcomes

Nursing defines a categorial structure as minimal set of categories and the relationships between them that are valid for representing concepts in terminological systems for a specified domain (see Fig. 8.19). The reference model for nursing diagnoses is defined by the graph depicted in Fig. 8.20. Nursing actions must have either a focus or content, and some actions contain both. The relationship between action content and targets is specified in Fig. 8.21.

These UML models represent the terminological models for the various forms of knowledge contained in any internationally standard nursing terminology as exemplified in ICNP.

The International Council of Nurses (ICN) is a federation of 129 national nurses associations. The International Classification for Nursing Practice (ICNP) is an official program of the ICN. The development and maintenance processes of the ICNP Program are used to increase the quality of ICNP. These include processes by which the ICNP was and continues to be developed, tested, distributed, and implemented worldwide, ICNP Version 1.0. The ICNP is a unified nursing language that facilitates cross-mapping among local existing terminologies. ICNP conforms to current terminology standards, for example, the ISO 18104 standard and HL7 terminology standards.

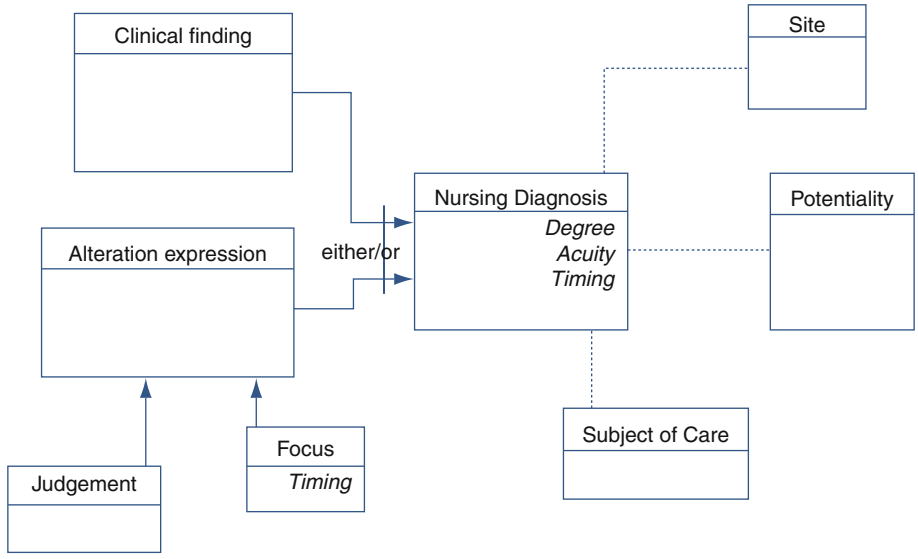
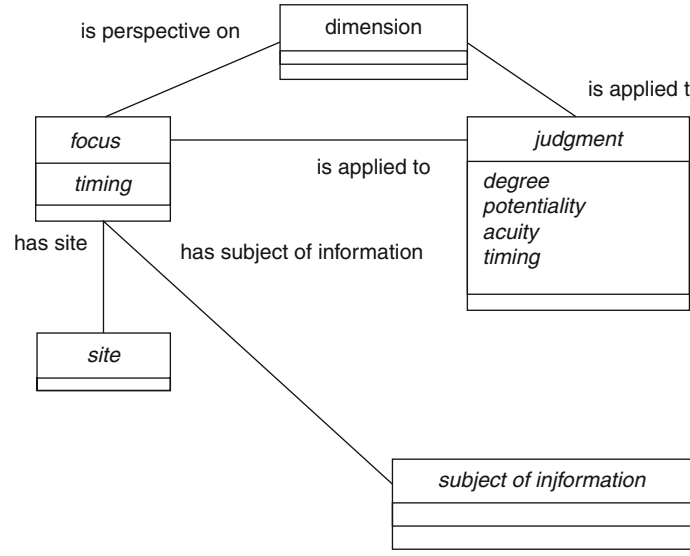


Fig. 8.19 Categorial structure for nursing diagnoses

Fig. 8.20 The reference model for nursing diagnoses



Reference terminology model for nursing diagnoses

The ICNP Alpha and Beta Versions document the history of concept validation and classification of nursing phenomena and interventions. The ICNP Beta 2 Version became a combinatorial terminology organized in two multiaxial hierarchies representing nursing phenomena and nursing actions. The ICNP Version 1.0, launched in 2005, changed the relatively straightforward multiaxial structure to a compositional terminology through the application of description logics instantiated using Web Ontology Language (OWL) within Protégé, a frame-based ontology development environment. ICNP Version 1.0 is also represented in a multihierarchical model (7-Hierarchies) for nurses

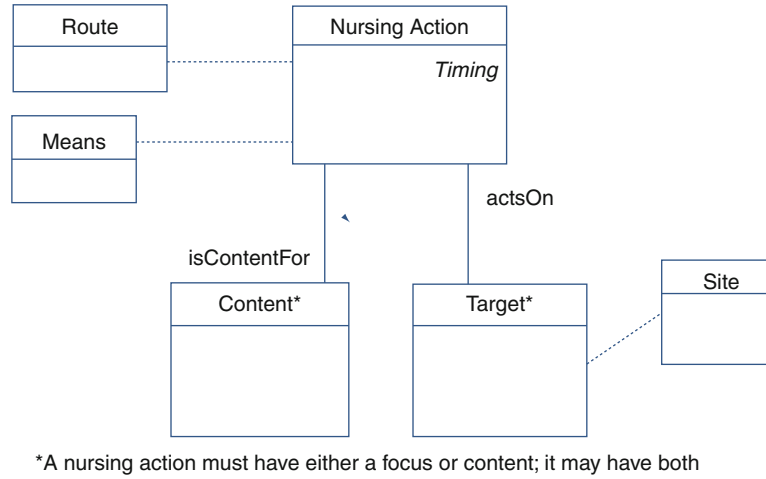


Fig. 8.21 Nursing action model

to compose nursing diagnosis, intervention, and outcome statements. Language translations and clinical information systems are utilized to make the ICNP Version 1.0 available to nurses at the point of healthcare delivery. ICNP data collected in healthcare environments provide standardized terminology for nursing that allows comparison of nursing practice across healthcare settings, specialties, and jurisdictional boundaries. The terminology is constructed to facilitate data-driven clinical decision making. ICNP can be employed in the development of guidelines and standards for best nursing practice toward optimal outcomes for patients, families, and their communities.

OBO and GO

The Open Biomedical Ontologies resource and its most well-known component the Gene Ontology are the most commonly used ontologies by molecular biologists. Open Biomedical Ontologies (OBO) is an effort to create controlled vocabularies for shared use across different biological and medical domains. OBO forms part of the resources of the US National Center for Biomedical Ontology, where it will form a central element of the NCBO's BioPortal. The OBO ontology library forms the basis of the OBO Foundry, a collaborative experiment involving a group of ontology developers who have agreed in advance to the adoption of a growing set of principles which they agree specify a set of best practices in ontology development. These principles are designed to foster the interoperability of ontologies, and to ensure a gradual improvement in the quality and formal rigor of ontologies, in order to meet the increasing information and knowledge requirements for the biomedical domain.

The Ontology Lookup Service (OLS) is a centralized query interface for ontology and controlled vocabulary retrieval. This service provides a standard interface and standard output for terminological data.

Gene Ontology Consortium

The goal of the Gene Ontology (GO) consortium is to produce a controlled vocabulary that can be applied to all organisms as knowledge regarding genes' and proteins' roles in cells is rapidly accumulating and evolving. GO provides three structured networks of defined terms to describe gene product attributes. GO holds the largest collection of molecular-biology-related terminological representation.

The Gene Ontology project provides an **ontology** of defined concepts representing gene product properties. The ontology covers the following three domains:

- **Cellular component**, the parts of a cell or its extracellular environment

- **Molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis
- **Biological process**, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms

Each GO concept within the ontology has a concept name, which may be a word or string of words; a unique alphanumeric identifier; a definition with cited sources; and a namespace indicating the domain to which it belongs. Concepts may also have synonyms, which are classed as being exactly equivalent to the concept name, broader, narrower, or related concepts; references to equivalent concepts in other databases; and comments on term meaning or usage.

The GO ontology is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain and sometimes to other domains. The GO vocabulary is designed to be species neutral and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms.

The GO ontology is not static, and additions, corrections, and alterations are suggested by, and solicited from, members of the research and annotation communities, as well as by those directly involved in the GO project. For example, an annotator may request a specific term to represent a metabolic pathway, or a section of the ontology may be revised with the help of community experts (e.g. The Glucocorticoid Metabolic Pathway). Suggested edits are reviewed by the ontology editors and after review are then implemented if and when they believe the additions to be appropriate.

The GO ontology file is freely available from http://gene-ontology.co.tv/ in a number of formats or can be accessed online using the GO browser AmiGO. The Gene Ontology project also provides downloadable mappings of its terms to other classification systems.

Sequence Ontology

The Sequence Ontology (SO) is a part of the Gene Ontology project, and the aim is to develop an ontology suitable for describing biological sequences. It is a joint effort by the genome annotation centers, including WormBase, the Berkeley Drosophila Genome Project, FlyBase, the Mouse Genome Informatics group, and the Sanger Institute.

Generic Model Organism Databases

The Generic Model Organism Database Project (GMOD) is a joint effort by the model system organism-related databases WormBase, FlyBase, MGI, SGD, Gramene, Rat Genome Database, EcoCyc, and TAIR to develop reusable components suitable for creating new biological database communities.

FGED

The Functional Genomics Data (FGED) Society is an international organization of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics experiments.

Plant Ontology

The Plant Ontology Consortium (POC) aims to develop, curate, and share structured controlled vocabularies (ontologies) that describe plant structures, their growth, and their developmental stages. The project aims to facilitate cross database inquiries by fostering the consistent use of these vocabularies in the annotation of tissue and their growth-stage-specific expression of genes, proteins, and phenotypes.

Phenoscape

Phenoscape is a project to develop a database of phenotype data for species across the Ostariophysi, a large group of teleost fish. The data are captured using annotations that combine concepts from an Anatomy Ontology, an accompanying Taxonomic Ontology, and quality concepts from the PATO

ontology of phenotype qualities. The anatomy ontology was developed from the zebrafish anatomy ontology developed by the Zebrafish Information Network.

OBO and Semantic Web

OBO and OWL Roundtrip Transformations

As a community effort, a standard common mapping has been created for lossless roundtrip transformations between Open Biomedical Ontologies (OBO) format and OWL. The research contains methodical examination of each of the constructs of OBO and a layer cake for OBO, similar to the Semantic Web stack.

Questions

1. The UMLS knowledge sources include:
   (a) Semantic Network
   (b) SPECIALIST Lexicon
   (c) The Metathesaurus
   (d) All of the above
   (e) b and c
2. The UMLS Metathesaurus includes:
   (a) One integrated terminology
   (b) A merger of multiple terminologies
   (c) A set of separate terminologies
   (d) A Semantic Network
3. The NLM stands for:
   (a) The Natural Language Machine
   (b) The National Library of Medicine
   (c) The Natural Library of Medicine
   (d) The Nation Library of Machine Learning
4. The UMLS terminologies can be found in:
   (a) The SPECIALIST Lexicon
   (b) The Semantic Network
   (c) The Metathesaurus
   (d) None of the above
5. The UMLS Semantic Network has how many semantic types?
   (a) 133
   (b) 103
   (c) 163
   (d) 193
6. In the UMLS, terminological information is linked by CUIs which stands for:
   (a) Contents unique identifiers
   (b) Contents unified by identifiers
   (c) Concept unique identifiers
   (d) Concept unified by identifiers
7. UMLS semantic types are classified as either:
   (a) Objects or classes
   (b) Entities or event
   (c) Entry or event
   (d) Entry or act
8. The difference between an SUI and an AUI is:
   (a) There are more SUIs than AUIs.
   (b) There are more AUIs than SUIs.
   (c) SUIs are unique to a string and a terminology.
   (d) AUIs are unique to a string and a terminology.
9. Synonyms in the UMLS are found in which file?
   (a) MRREL.RRF
   (b) MRCONSO.RRF
   (c) MRDEF.RRF
   (d) MRSTY.RRF
10. The UMLS is updated:
    (a) Once a year
    (b) Twice a year
    (c) Three times a year
    (d) Four times a year
11. The WHO Family of Classifications include all of the following except:
    (a) International Classification of Diseases (ICD)
    (b) International Classification of Function (ICF)
    (c) International Classification of Nursing Practice (ICNP)
    (d) International Classification of Health Interventions (ICHI)
12. In the United States, which version of the ICD is used to record a person's death?
    (a) ICD9
    (b) ICD9-CM
    (c) ICD10-AM
    (d) ICD10

13. Which version of ICD is used most commonly in research to record a patient's morbidity?
    (a) ICD9
    (b) ICD9-CM
    (c) ICD10-AM
    (d) ICD10
14. ICD9-CM is maintained by:
    (a) The CDC
    (b) The NLM
    (c) The WHO
    (d) The IHTSDO
15. Error rates in ICD9-CM code assignments are best represented by which rate?
    (a) 2%
    (b) 12%
    (c) 22%
    (d) 32%
16. Which of the following statements are true?
    (a) SNOMED CT has more concepts than ICD9-CM.
    (b) ICD9-CM has more concepts than SNOMED-CT.
    (c) ICD9-CM has been used in more research projects than SNOMED CT.
    (d) a and c
    (e) b and c
17. The following terminologies have a description logic basis except:
    (a) SNOMED CT
    (b) ICNP
    (c) NDF-RT
    (d) LOINC
18. The following terminologies have precoordinated concepts:
    (a) LOINC
    (b) RxNorm
    (c) ICD9-CM
    (d) All of the above
19. Which terminologies have been used as a reference terminology?
    (a) SNOMED CT
    (b) ICF
    (c) NDF-RT
    (d) LOINC
    (e) a and c
20. SNOMED CT is maintained by the:
    (a) WHO
    (b) IHTSDO
    (c) NLM
    (d) CDC
21. The form of compositional expressions in SNOMED CT is governed by:
    (a) Common sense
    (b) The Description Logic Handbook
    (c) The SNOMED CT style guide
    (d) The IHTSO Handbook
22. In SNOMED CT, the description logic definitions are used:
    (a) To serve as a formal definition for the concept
    (b) To find a concepts place in the SNOMED CT hierarchy
    (c) To find conflicts in the terminology
    (d) All of the above
23. LOINC was originally created to represent:
    (a) Radiological knowledge
    (b) Laboratory knowledge
    (c) Pathological knowledge
    (d) Clinical practice knowledge
24. LOINC is maintained by:
    (a) WHO
    (b) IHTSDO
    (c) NLM
    (d) CDC
    (e) Regenstrief Institute
25. The UMLS contains:
    (a) LOINC
    (b) SNOMED CT
    (c) ICD9-CM
    (d) All of the above
26. Which of the following is/are true?
    (a) NDF-RT and RxNorm have some of the same concepts.
    (b) NDF-RT and RxNorm have all the same concepts.
    (c) NDF-RT and RxNorm have the same hierarchical information.
    (d) NDF-RT and RxNorm are both maintained by the NLM.
27. Which of the following is/are true?
    (a) LOINC concepts can be postcoordinated.
    (b) LOINC analytes are multipart names.

(c) LOINC components are compositional.
(d) LOINC clinical concepts are limited to laboratory-related clinical findings.

28. The five-digit subdivisions were developed for:
(a) ICD0
(b) ICD5
(c) ICD9
(d) ICD10

29. SNOMED CT was formed as a merger of:
(a) ICD9-CM and LOINC
(b) CAPs terminology effort and the NHS terminology effort
(c) Read Codes v2 and SNOMED International
(d) Clinical Terms v3 and SNOMED II

30. SNOMED CT has its headquarters in:
(a) United States
(b) Great Britain
(c) Denmark
(d) France

31. OCD in RxNorm stands for:
(a) Obsessive compulsive disorder
(b) Overly crowded directory
(c) Obsolete clinical drug
(d) Open component drug

32. SNOMED CT's coverage rate for clinical problems using only precoordination is:
(a) 61.4%
(b) 51.4%
(c) 41.4%
(d) 31.4%

33. SNOMED CT's coverage rate for clinical problems using postcoordination is:
(a) 100%
(b) 92.3%
(c) 82.3%
(d) 72.3%

34. SNOMED CT is maintained by:
(a) IHTSDO
(b) IMIA
(c) NLM
(d) CAP

35. Which statements are true regarding SNOMED CT?
(a) SNOMED CT covers all nursing content.
(b) SNOMED CT covers all clinical drugs.
(c) SNOMED CT has incorporated lab LOINC into its hierarchies.
(d) SNOMED CT has incorporated all of ICD10 into its hierarchies.

36. The Therapeutic Intent hierarchy is from:
(a) SNOMED CT
(b) NDF-RT
(c) RxNorm
(d) LOINC

37. Nursing terminologies include all the following except the:
(a) ICNP
(b) NANDA
(c) ICHI
(d) LOINC

38. In ICNP, an attribute of focus is:
(a) Concentration
(b) Adjustment
(c) Focal length
(d) Timing

39. ICNP is distributed in:
(a) Common logic
(b) OWL
(c) LOINC format
(d) SNOMED CT format

40. Nursing terminologies have led the development of:
(a) Concept-based terminologies
(b) Description logics
(c) Goal statements
(d) Mortality coding

41. When was CPT first published?
(a) 1956
(b) 1960
(c) 1966
(d) 1976

42. How many categories exist for CPT codes?
(a) 2
(b) 3
(c) 4
(d) 5

43. How many sections are the Category I CPT codes subdivided into?
(a) 2
(b) 6
(c) 4
(d) 5

44. The assignment of a procedure or service to a specific CPT section restricts its use by specific healthcare specialties. True or False?
    (a) True
    (b) False
45. CPT Category III codes are permanent tracking codes used to report the supply of drugs. True or False?
    (a) True
    (b) False

References

1. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med. 1993;32(4):281–91.
2. Farr W. Regarding the Cullenian system of 1785, first annual report of the Registrar-General of births, deaths, and marriages in England. London; 1839 p. 99.
3. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS, for the Canon Group. Toward a medical-concept representation language. JAMIA. 1994;1:207–17.
4. http://loinc.org/.
5. World Health Organization. International classification of impairments, disabilities, and handicaps. A manual of classification relating to the consequences of disease. Geneva: World Health Organization; 1980.
6. Musen MA, Wieckert KE, Miller ET, Campbell KE, Fagan LM. Development of a controlled medical terminology: knowledge acquisition and knowledge representation. Methods Inf Med. 1995;34(1–2):85–95.
7. Elkin PL, Brown SH, Husser C, Bauer BA, Wahner-Roedler D, Rosenbloom ST. An evaluation of the content coverage of SNOMED-CT for clinical problem lists. Mayo Clin Proc. 2006;81(6):741–8.
8. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. A model for evaluating interface terminologies. J Am Med Inform Assoc. 2008;15(1):65–76. Epub 2007 Oct 18.
9. Brown SH, Husser C, Wahner-Roedler D, Bailey S, Nugent L, Porter K, Bauer BA, Elkin PL. Using SNOMED-CT as a reference terminology to cross map two highly pre-coordinated classification systems". Medinfo. 2007;12(Pt 1):636–9.
10. Percy C, Van Holten V, Muir C, editors. International classification of diseases for oncology (ICD-O). 2nd ed. Geneva: World Health Organization; 1990. www.who.int/classifications/icd/ICD-10_2nd_ed_volume2.pdf
11. College of American Pathologists. Systematized nomenclature of medicine (SNOMED). Chicago: College of American Pathologists; 1976.
12. Percy C, Berg JW, Thomas LB, editors. Manual of tumor nomenclature and coding (MOTNAC). New York: American Cancer Society; 1968.
13. College of American Pathologists. Systematized nomenclature of pathology (SNOP). Chicago: College of American Pathologists; 1965.
14. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva: World Health Organization; 1992.
15. World Health Organization. International classification of procedures in medicine (ICPM), vol. 1 and 2. Geneva: World Health Organization; 1978.
16. International Nomenclature of Diseases. APA (6th ed.) Council for International Organizations of Medical Sciences., & World Health Organization. (1900). International nomenclature of diseases. Geneva: CIOMS.
17. REGISTRAR General. Sixteenth annual report. London: Registrar General of England and Wales; 1856. p. 73.
18. Knibbs GH. The international classification of disease and causes of death and its revision. Med J Aust. 1929;1:2–12.
19. Greenwood M. Medical statistics from Graunt to Farr. Cambridge University Press: Cambridge; 1948.
20. Registrar General of England and Wales. First annual report. London: Registrar General of England and Wales; 1839. p. 99.
21. Bulletin of the Institute of International Statistics. 1900;12:280.
22. Roesle E. Essai d'une statistique comparative de la morbidité devant servir à établir les listes spéciales des causes de morbidité. Geneva: League of Nations Health Organization; 1928 (document C.H. 730).
23. International Statistical Institute. Nomenclatures internationales de causes de décès. The Hague: International Statistical Institute; 1940.
24. Medical Research Council, Committee on Hospital Morbidity Statistics. A provisional classification of diseases and injuries for use in compiling morbidity statistics. London: Her Majesty's Stationery Office; 1944 (Special Report Series No. 248).
25. http://www.nlm.nih.gov/research/umls/rxnorm/.
26. http://www.icn.ch/pillarsprograms/about-icnpr/.
27. Bertillon J. Classification of the causes of death (abstract). http://archive.org/details/bertillonclassif00amer In: Transactions of the 15th international congress on hygiene demography, Washington, DC; 1912.
28. Nelson SJ, Brown SH, Erlbaum MS, Olson N, Powell T, Carlsen B, Carter J, Tuttle MS, Hole WT. A semantic normal form for clinical drugs in the UMLS: early experiences with the VANDF. Proc AMIA Symp. 2002:557–561.

Terminological Systems

9

Peter L. Elkin and Mark Samuel Tuttle

This chapter describes some successful terminological systems. It is not possible to describe all terminological efforts as there are too many; however, I will cover examples of the most commonly employed types of software that are used to build, maintain, distribute, use, or employ terminological systems. Some categories are so large that I will provide multiple examples. We have used some of the publically available materials and statements regarding the descriptions of vended systems used as examples of types of terminological systems in this chapter. The categories and systems that will be discussed in their historical order of appearance are:

Terminology or ontology authoring and maintenance environments

- Apelon TDE
- Protégé
- IHTSDO Workbench

Natural language processing systems

- MedLEE
- iNLP
- MetaMap
- NEGEX/OpenNLP

Terminology servers and services

- DTS
- iNLP server
- CTS II

Secondary use of clinical data

- Opticode
- DIEL
- Marker Discovery
- Drug labels
- Clinical decision support systems
- QMR
- DXplain
- Iliad

P.L. Elkin, M.D., MACP, FACMI (✉)
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

M.S. Tuttle, AB, BE, FACMI
Apelon, Ridgefield, CT, USA

Terminology or Ontology Authoring and Maintenance Environments

Apelon TDE

Apelon Corporation's terminology development environment (TDE) was one of the earliest description logic based terminology authoring environments. It was used for the initial construction of SNOMED RT and SNOMED CT. The TDE was also employed in the Convergent Medical Terminology project at Kaiser Permanente [1].

Apelon's terminology development environment (TDE) is developed to facilitate the creation, maintenance, and evolution of structured terminologies and ontologies. The TDE has helped to improve the quality and efficiency of the complex, people-intensive, and time-consuming task of developing formal, structured terminologies for customers like the American Medical Association, Centers for Disease Control, College of American Pathologists, ECRI, Kaiser

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_9,

Permanente, Motorola, National Cancer Institute, and the US Department of Veterans Affairs.

Apelon's TDE combines a powerful, customizable, GUI-based authoring/editing application with a terminology engine based on description logic (based on K-Rep). Advanced features of the TDE include version management and control, workflow and conflict resolution for distributed authoring, and flexible methods for exporting and exchanging terminology data.

For organizations needing to create and evolve critical terminology assets, TDE provides a solution that Apelon states will reduce maintenance costs, increase terminology quality, and improve results.

Protégé

Protégé is a free, open-source ontology editor and knowledge-base framework [2]. It is a free, open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-based applications with ontologies. At its core, Protégé implements a rich set of knowledge-modeling structures and actions that support the creation, visualization, and manipulation of ontologies in various representation formats. Protégé can be customized to provide domain-friendly support for creating knowledge models and entering data [3]. Further, Protégé can be extended by way of a plug-in architecture and a Java-based application programming interface (API) for building knowledge-based tools and applications.

An ontology describes the concepts and relationships that are important in a particular domain, providing a vocabulary for that domain as well as a computerized specification of the meaning of terms used in the vocabulary. Ontologies range from taxonomies and classifications, database schemas, to fully axiomatized theories. In recent years, ontologies have been adopted in many business and scientific communities as a way to share, reuse, and process domain knowledge. Ontologies are now central to many applications such as scientific knowledge portals, information management and integration systems, electronic commerce, and semantic web services.

The Protégé platform supports two main ways of modeling ontologies:

- The **Protégé-Frames** editor enables users to build and populate ontologies that are *frame-based*, in accordance with the Open Knowledge Base Connectivity protocol (OKBC). In this model, an ontology consists of a set of classes organized in a subsumption hierarchy to represent a domain's salient concepts, a set of slots associated to classes to describe their properties and relationships, and a set of instances of those classes – individual exemplars of the concepts that hold specific values for their properties.
- The **Protégé-OWL** editor enables users to build ontologies for the *Semantic Web*, in particular in the W3C's Web Ontology Language (see the OWL Web Ontology Language Guide).

The Protégé-OWL editor is an extension of Protégé that supports the Web Ontology Language (OWL). OWL is the most recent development in standard ontology languages, endorsed by the World Wide Web Consortium (W3C) to promote the *Semantic Web* vision. "An OWL ontology may include descriptions of classes, properties and their instances. Given such an ontology, the OWL formal semantics specifies how to derive its logical consequences, i.e. facts not literally present in the ontology, but entailed by the semantics. These entailments may be based on a single document or multiple distributed documents that have been combined using defined OWL mechanisms," from the Web Ontology Language Guide [4].

The Protégé-OWL editor enables users to:

- Load and save OWL and RDF ontologies
- Edit and visualize classes, properties, and Semantic Web Rule Language (SWRL) rules
- Define logical class characteristics as OWL expressions
- Execute reasoners such as description logic classifiers
- Edit OWL individuals for Semantic Web markup

Protégé-OWL's flexible architecture makes it easy to configure and extend the tool. Protégé-

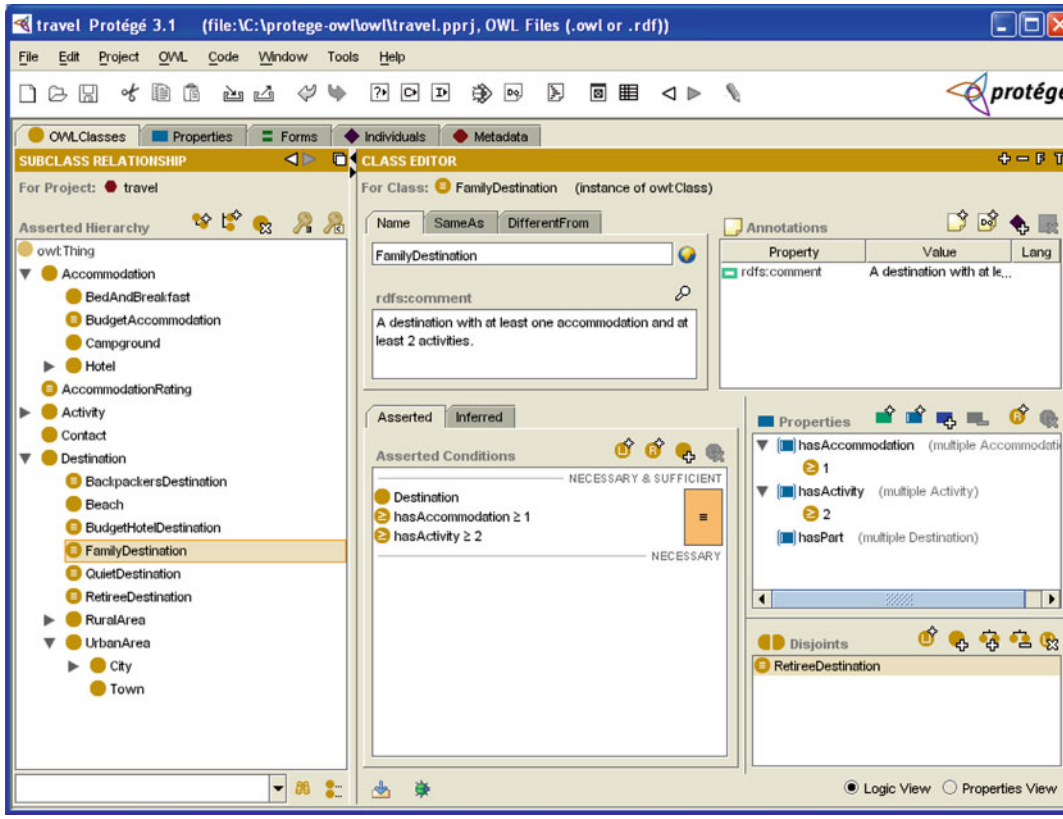


Fig. 9.1 Protégé OWL based representation of information related to a family vacation

OWL is tightly integrated with Jena and has an open-source Java API for the development of custom-tailored user interface components or arbitrary Semantic Web services (see Fig. 9.1).

IHTSDO Workbench

The IHTSDO Workbench provides two software frameworks for terminology-focused applications: one for build-process automation and another for interactive development environments [5]. These frameworks provide application skeletons that can be customized by an application developer to meet unique needs of their end users.

The *first framework* – the IHTSDO Build Process Automation (BPA) framework – is based on the Maven tool for building and managing any Java-based project (The Apache Software Foundation, 2008). This BPA framework automates a build and management process that encourages use of industry best practices. The IHTSDO BPA framework builds on the Maven foundation by providing terminology-specific functions to manage, process, and report terminology and classification data dependencies within JAR files, thereby providing a uniform framework for managing software and terminology dependencies.

The *second framework* – the IHTSDO Interactive Development Environment (IDE) framework – provides for high-performance end-user applications and uses a Java Swing-based framework that is easily extended and scripted using plain-old java objects.

The International Healthcare Terminology Standards Development Organization (IHTSDO) is a not-for-profit association that develops and

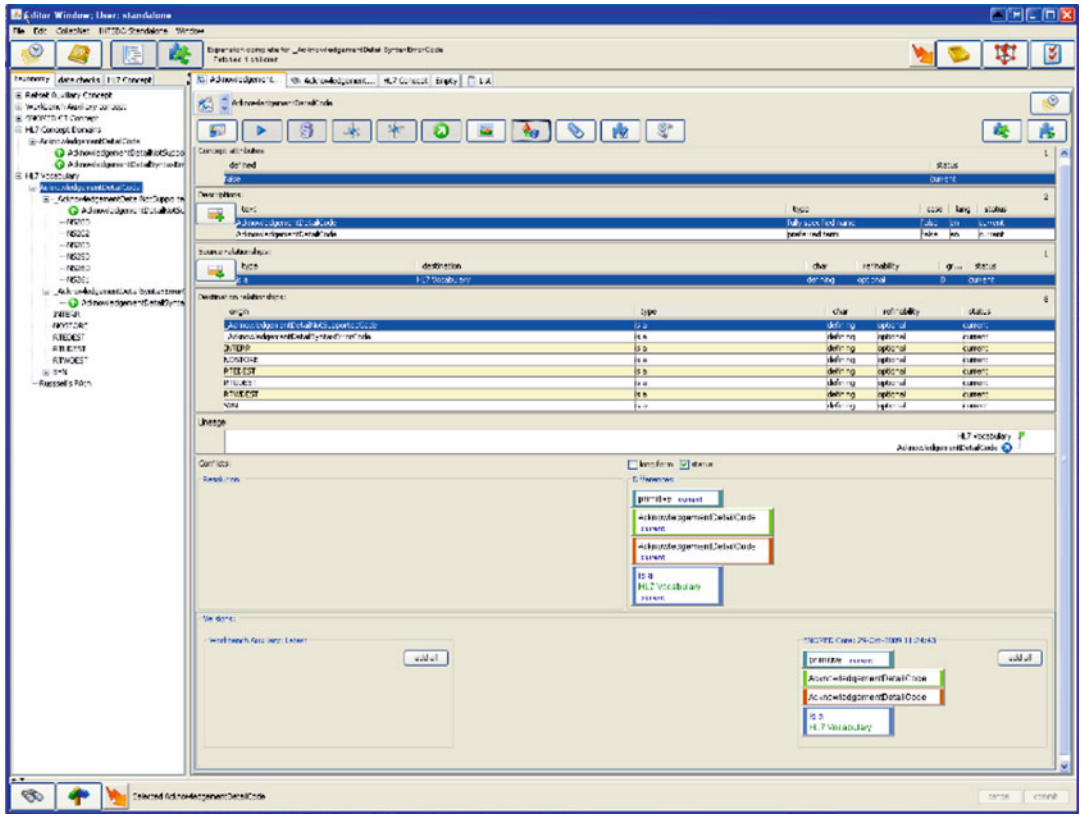


Fig. 9.2 IHTSDO Workbench modeling interface example

promotes use of SNOMED CT to support safe and effective health information exchange.

SNOMED CT is a clinical terminology and is considered to be one of the most comprehensive, multilingual healthcare terminologies in the world.

SNOMED CT consists of approximately 291,000 active concepts as of July 2010, arranged in a hierarchy, connected by relationships. SNOMED CT semantics are based on description logic.

In January 2009, the IHTSDO procured a workbench to maintain SNOMED CT, as part of an Open Health Tools (OHT) Charter project. The workbench contains the following functionality:

- Terminology life cycle management
- Automated workflow
- Searching, browsing, and editing support
- Support for reference sets
- Support for cross-mapping to other terminologies and code sets
- Support for classification
- Build process automation
- Change management and conflict resolution
- Collaboration tools
- Project management and support tools
- Lexical support

In December 2009, IHTSDO made the source code for the workbench open source under an Apache-2 license.

The workbench allows users to browse and author terminological content (see Fig. 9.2). It has a description logic classifier built into the system to assist users with the classification task.

The process portion of the workbench allows users to define workflows (see Fig. 9.3). This is important for terminological review and mapping tasks. The workbench has been used to map ICD10 to SNOMED CT.

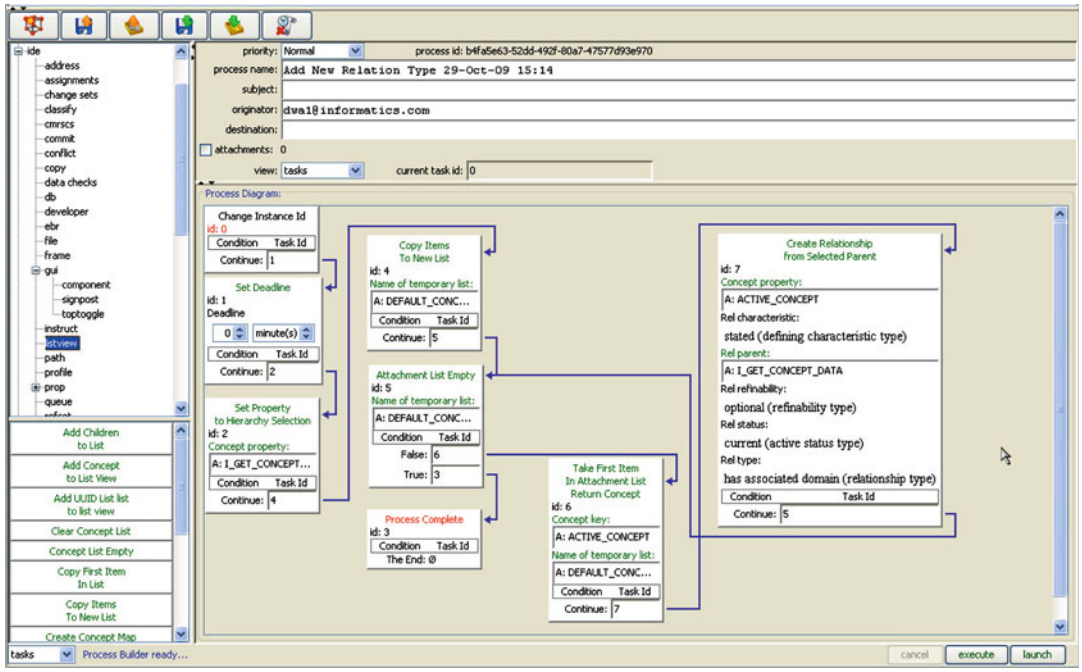


Fig. 9.3 IHTSDO Workbench workflow interface example

Natural Language Processing Systems

MedLEE

The Medical Language Extraction and Encoding System (MedLEE) was developed by Carol Friedman, Ph.D., of Columbia University [6]. The system is marketed by NLP International.

MedLEE is a text processor that extracts and structures clinical information from textual material including radiology reports and translates the information to terms in a controlled vocabulary, such as the UMLS or SNOMED. Clinical information can then be accessed by further automated methods. Although the processor has been applied to the domains of radiology, discharge summaries, sign-out notes, pathology reports, electrocardiogram reports, and echocardiogram reports, the design is extensible so that it can readily be ported to other clinical domains. MedLEE generates XML output and has the ability to incorporate local terminology.

MedLEE is a crucial component of an innovative knowledge management tool. The system allows for the automation and simplification of decision making when integrated into a clinical information system. MedLEE can be configured for applications other than medical reports.

In recent clinical use, MedLEE has been used in a patient safety program to detect a broad range of medical events: the 45 patient-specific events defined in the New York Patient Occurrence Reporting and Tracking System. The system achieved very high performance, enabling broad screening for medical events.

MedLEE has also been applied to improve safety through intervention. For example, MedLEE detects patients at high risk for having active tuberculosis and recommends respiratory isolation. Its use has reduced the rate of missed tuberculosis by almost one half at New York-Presbyterian Hospital.

The output of MedLEE includes a set of matched classes and qualifiers providing information about the context of the matched

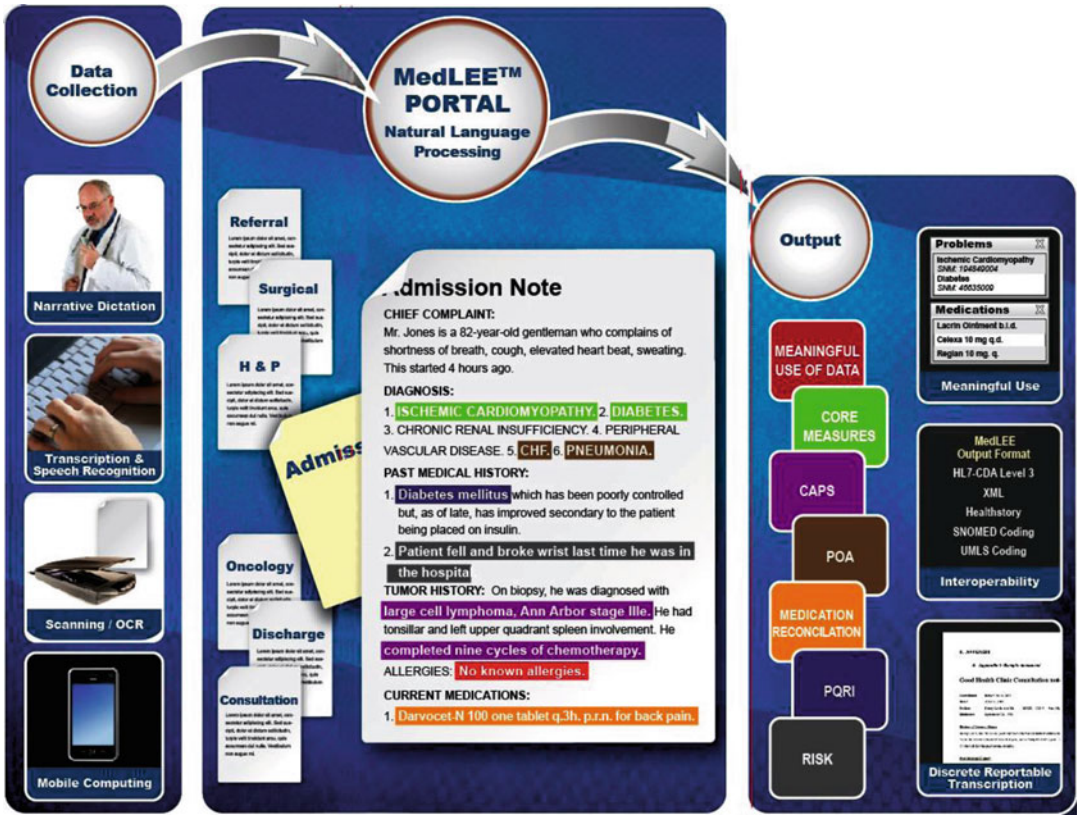


Fig. 9.4 MedLEE Portal example

information from either SNOMED CT or the UMLS (see Fig. 9.4).

iNLP

The intelligent natural language processor (iNLP) was developed by Dr. Peter Elkin, and its earlier incarnations have been used by the Veterans Administration, the Centers for Disease Control and Prevention, Johns Hopkins University, Vanderbilt, University of Pennsylvania, Mount Sinai, and other healthcare institutions nationally and internationally.

A significant portion of health data is locked and noncomputable in the form of free text. The multithreaded clinical vocabulary server (iNLP) is the most accurate natural language processor in existence with sensitivities of 99.7% and positive predictive value of 99.8%. This capability provides public health with advanced access to data needed to fuel biosurveillance/situational awareness, chronic disease management, and clinical decision support. This core will provide vocabulary services to each of the individual projects in this program project grant.

Laboratory of Biomedical Informatics (LBI) Experience

The clinical problem list is an important source of information about a patient. It is a regulatory requirement [7], as well as a recommended method of communicating the state of a patient's health [8]. Given physicians' dislike for structured data entry, an attractive feature of an EHR-S is the ability to link the physician's free-text-entered problem with an underlying structured

vocabulary. The author of this chapter and colleagues at the LBI researchers have worked for the past several years in providing that feature, resulting in the creation of the multithreaded clinical vocabulary server. Figures 9.5, 9.6, and 9.7 display the output of the system that addresses not only the problem list but the entire clinical record.

In this study, we compare the sensitivity, specificity, positive likelihood ratio, and positive predictive value of SNOMED-CT in providing content coverage for the information stored in patient problem lists [9].

We selected the 5,000 most common nonduplicated (unique text strings) from the Mayo Master Sheet Index associated with episodes of care from both the inpatient and the outpatient setting. Each record had associated with it, the free text final diagnoses from the Master Sheet Index at the Mayo Clinic. The free text diagnoses were coded by two physicians (disagreements were addressed by an expert clinician/terminologist) as to whether the terminology (SNOMED CT) was able to represent the problem. The free text entries were also automatically coded using the iNLP server. Reviewers also had available to them a SNOMED CT browser to allow them to look up any nonexact matches. Each problem was compared with the gold standard created by the expert indexers.

SNOMED CT had coverage of 92.3% for 4,996 common problem statements, which served as the test set for this study. SNOMED CT correctly identified 4,568 terms (48.9% required a compositional expression to exactly represent the concept), 36 terms were not felt to be sensible expressions or were misspelled and were not matched by SNOMED CT, 9 were felt not to be sensible but were matched by SNOMED CT, and 383 terms were felt to be sensible but did not match exactly to SNOMED CT. In this study, SNOMED CT had sensitivity (recall) of 92.3%, a specificity of 80.0%, and a positive predictive value (precision) of 99.8%. After correction for missing or erroneous synonymy in SNOMED CT, the iNLP engine had sensitivity (recall) of 99.7%, a specificity of 97.9%, and a positive predictive value (precision) of 99.8%. The positive likelihood ratio (positive likelihood ratio = sensitivity/(1 – specificity). A value of 1 means no discrimination (ability to cover a free-text term); values over 20 are excellent the positive likelihood ratio for the iNLP rose to 47.5, and the negative predictive value was 97.0%. The interrater reliability for their judgment regarding SNOMED CT was 91.8% with a Kappa of 0.49 and for the iNLP system was 94.3% with a Kappa score of 0.79.

We concluded that SNOMED CT has good coverage of the terms used commonly in medical problem lists. Improvements to synonymy and

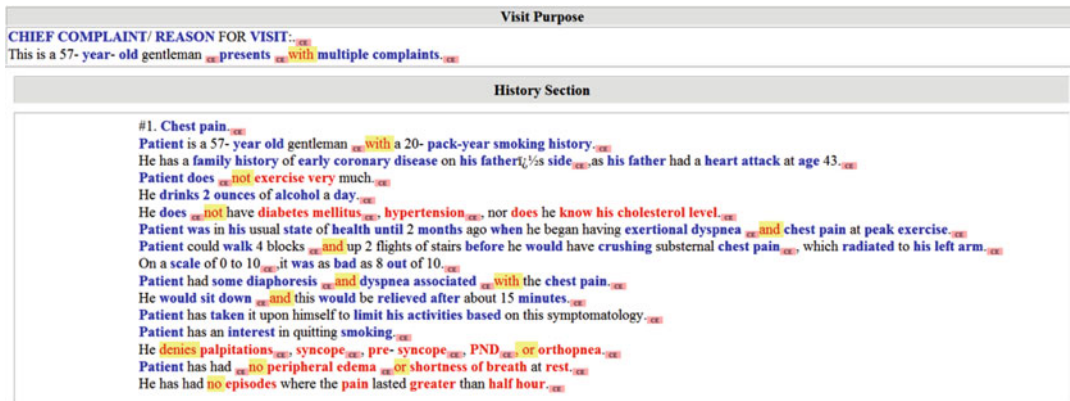


Fig. 9.5 Indexed history by the intelligent natural language processor (iNLP). All concepts are coded using SNOMED CT. Blue concepts are positive assertions, red concepts are negative assertions, and green concepts are uncertain assertions

	Exam Section
Vital Signs	PHYSICAL EXAMINATION: HEIGHT: 190 cm WEIGHT: 110 kg. TEMP: 36.3 C HEAD CIRCUMFRANCE: PULSE: 84 RHYTHM: Regular. SBP: 138 DBP 82 Position Date / Time:
Eyes	Eyes Non- icteric. Pupils were equal and reactive to light and accomodation.
ENT	ENT Ears are clear. Oral cavity, or al pharynx is clear.
Thyroid	Thyroid Neck is supple without nodes or masses. Thyroid is within normal limits.
Vessels	Vessels Carotid Arteries are 2+ without bruits.
Heart	Heart Normal S1, normal S2; without murmurs, gallops, rubs, or clicks.
Lungs	Lungs Clear without wheezing, rales, rhonchi, or rubs.
Abdomen	Abdomen Soft, flat, non- tender, normal active bowel sounds without hepatosplenomegaly or masses or bruits.
Rectum	Rectum brown stool at the verge, no other masses.
Genitalia	Genitalia Within normal limits. He had no lesions. He had no testicular masses.
Extremities	Extremities Without clubbing, cyanosis, or edema.
Gait	Gait Within normal limits.
Neuro	Neuro Cranial nerves 2 through 12 were intact. Visual fields were within normal limits. Sensation was intact and bilaterally symmetric. Motor was 5+/5+ bilaterally symmetric. Deep tendon reflexes were 2+/2+ and were symmetric bilaterally. Romberg was normal. Cerebellar signs were absent. Babinski was down going bilaterally.

Fig. 9.6 Indexed physical examination utilizing the iNLP

the addition of missing modifiers would lead to the greatest return on investment toward improved coverage of common problem statements. Compositional expressions are required to exactly represent a significant portion of common problem statements.

iNLP Technical Summary

The iNLP system is a compendium of advanced indexing tools. The system indexes source materials using a concept-based indexing schema. The iNLP system is the first multilingual natural language processor which reads both English and French and is currently implementing Spanish and Chinese. This underlying indexing schema is terminology independent but architected to take advantage of the robust ontology of medical concepts available in the SNOMED CT terminology. The development of the iNLP has been based on more than 23 years of research and has been an ongoing development effort for more than three years, which has yielded a product of considerable accuracy and flexibility. The iNLP is based on a robust underlying terminological model and a component architecture, which uses industry

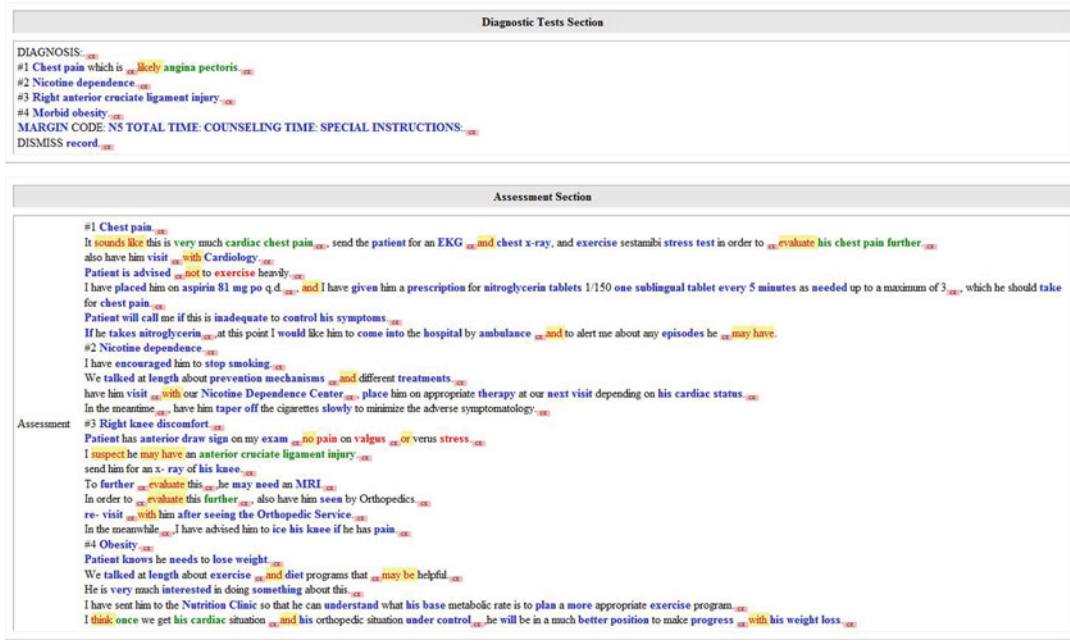


Fig. 9.7 Impression section of the record indexed by the iNLP

standard technologies (relational databases, an RMI server, Enterprise Java Bean middleware objects, Java Client Interfaces, and Java Server Pages in a browser environment). The software has been extensively tested in LBI's Usability Laboratory and has been presented at the *American Medical Informatics Association's Fall Symposium*. This four-tier architecture has improved our system throughput by two orders of magnitude. The Java Remote Method Invocation (RMI) servers are specific high-performance dual processor machines that cache Java objects in memory, greatly decreasing the number of queries to the database and hence greatly increasing system performance (see Fig. 9.8). This makes practical, the processing of millions of records during the study period. More recently, the iNLP system was written in the .NET architecture using the C# programming language. This system uses a .NET cash and works seamlessly with SQL Server's cash to provide a threefold improvement in throughput.

Vocabularies such as SNOMED CT, the UMLS, or one of several other vocabulary efforts have common characteristics that are helpful to understand within the scope of iNLP. A vocabulary is essentially a set of concepts that are identified by a unique identifier and described by terms and relationships. For each concept, there exists one or more terms that belong to that concept. For example, in SNOMED CT, the concept "Myocardial Infarction" is identified by "22298006" and contains the terms "heart attack," "infarction of heart," and "cardiac infarction." The concept is also described by its relationships to other concepts. These relationships commonly include hierarchies built on parent/child relationships but often extend to many other types of relationships such as morphology, topology, and etiology.

The iNLP system includes a terminology browser which is depicted below.

Practice

*e*Quality

All mainstream safety and quality programs require practitioners to utilize secondary data and human accounts to discover root causes of what

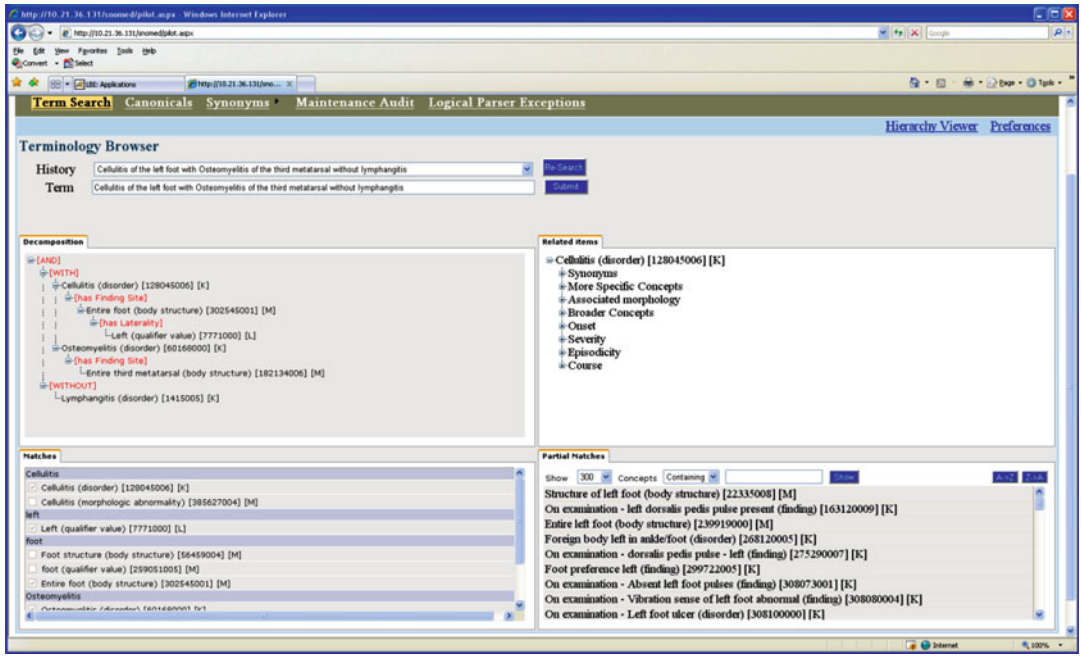


Fig. 9.8 iNLP Terminology Browser

is done well and not so well during an event. Healthcare data, with its rich objective data (lab results, time stamps, flow charts, etc.) and descriptive data (physician notes, impressions, and patient accounts), have the potential to deliver a clearer picture of the actual circumstances surrounding both good and poor quality than your typical manufacturing or service provider. So why is healthcare considered such a poor quality product, and what are the barriers to improving healthcare quality?

The United States Agency for Healthcare Research and Quality (AHRQ) lists five reasons for poor quality in healthcare as follows:

1. Variation in services
2. Underuse of services
3. Overuse of services
4. Misuse of services
5. Disparities in quality

The use of secondary data in *e*Quality can help to address each of these points.

Variation in services – While we understand that variation in services occurs in practice, the industry has done very little to assure patients and the practice that best practice is utilized. This is not necessarily a physician problem. Just to keep up, the average practitioner must read dozens of journals per week. This is logistically impossible. This can lead to the unsafe practice of medicine.

The use of secondary data can allow for retrospective or almost real-time review of clinical protocols to both suggest and aid the physician in the treatment plan. Currently there are two hurdles to this approach. The first is that there is little or no integration of best practices in most clinical settings. Second, there were and there is no governing body dictating the one best practice.

Secondary data, longitudinally applied, tracked, and evaluated, offer great promise to the evaluation of best practices. The President of the United States has proposed panels to evaluate cost-effectiveness (where do we get the best results per dollar spent), but has met resistance by some who see this as rationing. Understanding what does and does not work through the use of secondary data allows for sound decision making is a primary driver for consideration as EHRs are planned and executed and related directly to AHRQ's list of reasons for poor quality.

Underuse of services – Secondary data allow the clinical provider to obtain evidence of underuse of services by omission. If one expects certain immunizations for a child in the first several years of life and secondary data do not suggest that it has been done, underuse issues can be more easily identified than how it is done today, which is typically by a school nurse as the child enters kindergarten. This is also true for cholesterol, mammograms, and colonoscopies.

Overuse of services – Defensive medicine, poor communication, and variations in practice contribute to the overuse of medical services. This could include the wrong prescriptions, excessive prescriptions, too many tests, a non-indicated test, or an improper dosing. Secondary data allow analysts to statistically analyze practice patterns for outliers that can help providers control the overuse of services. For example, if analysis of data indicates a high rate of c-sections in an institution, it is incumbent on the facility to act, but if secondary data indicate that a community has high rates of c-sections, it is someone else's duty to act.

Misuse of services – Adverse events are well-documented in healthcare. When an event happens, there is usually an investigation (assuming it is reported), and that investigation typically identifies the cause of the event. What secondary data allow one to do is to analyze groups of adverse events to discover things that analysis of a single event does not. Alternatively, secondary data allow us to develop profiles that potentially capture events that are incurred, but not reported (IBNR). An example might include triggers or other leading or trailing indicators that may indicate a miss or what might be considered a minor event in one case, which over a period of time and a number of cases might be considered seminal.

Disparities in quality – The use of secondary data could be a powerful tool in the exposure of disparities. Analysis of healthcare delivery between races, socioeconomic classes, and gender would be greatly enhanced and potentially be eliminated if secondary data, available across the healthcare continuum, were available to be statistically analyzed.

Finally, in the pursuit of eQuality, clinicians need depth of information and not just width. Specifically, a lot of secondary data are no substitute for knowledge and information. So the strategic discussions necessary as the USA rolls out its EHR programs are severalfold including planning for the long-term disposition of that data (environmental and preventive health), strategically planning for anticipated future needs (genomic data), and integration of best practices into existing systems (what is "alertable").

The eQuality initiatives at the VA by Brown et al. serve as excellent examples of the use of secondary data toward improved and safer patient care [10, 11].

Education

Multimedia linked to clinical cases have been organized into digital education libraries. Linkage of case-based teaching tools using ontological encoding of the clinical data using the same standards used to index the clinical data (e.g., SNOMED CT, LOINC, and RxNorm) can serve to link teaching materials to clinical cases. This can and will enable continuous quality improvement to the care process and continuing medical education at the point of care. In the example below, we encode case-based data and answers in a case-based teaching web-based tool (see Figs. 9.9 and 9.10).

Research

Secondary use of clinical data for research is one of more well thought out areas of application. For many years, we have used patient records as a data source for human abstraction of clinical research data. With the advent of electronic health record (EHR) data, we can now make use of computable EHR data that can perform retrospective research studies more rapidly and lower the activation energy necessary to ask the next important question using electronic studies (*e*Studies). Barriers to these eStudies include: the lack of interoperable data between and among practices,

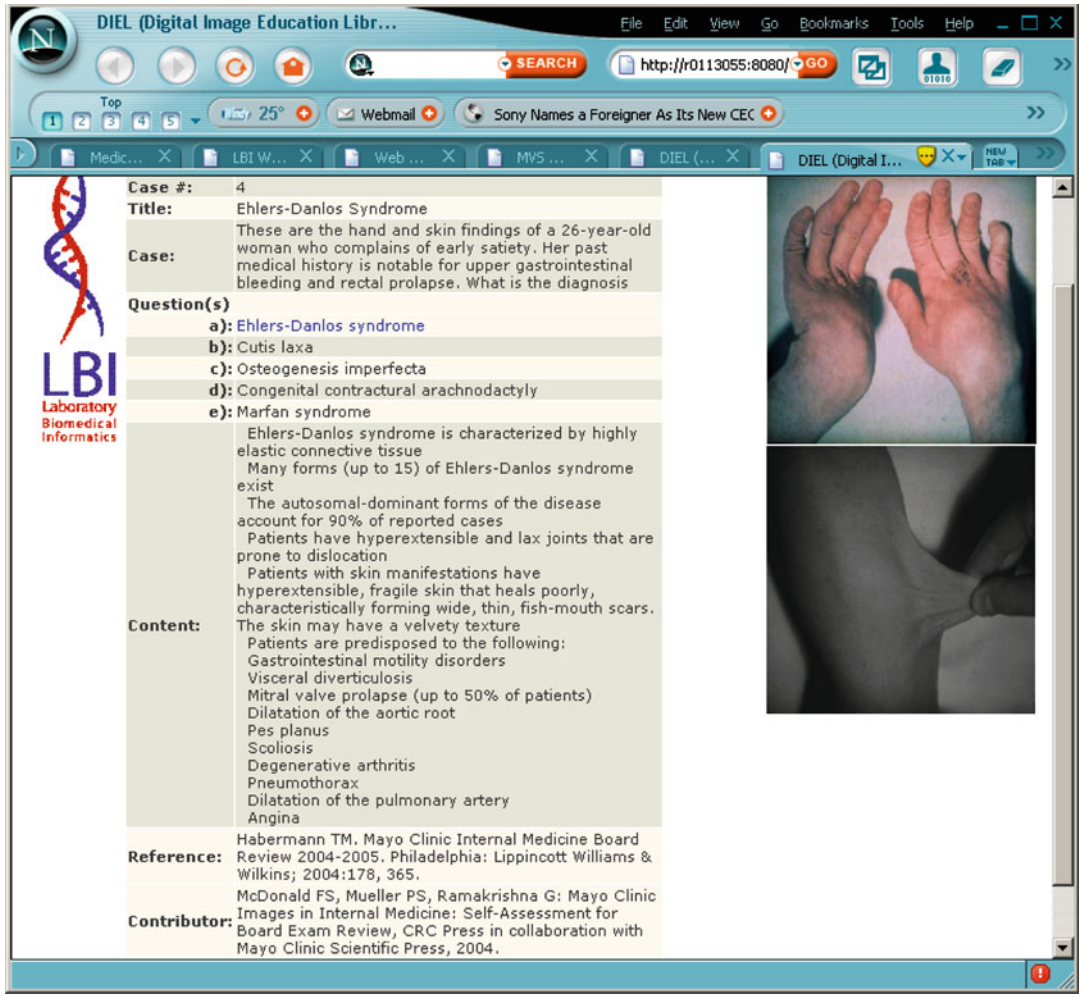


Fig. 9.9 A case of Ehlers–Danlos syndrome with teaching content all indexed using SNOMED CT

the lack of computable definitions of measures, the lack of training of health professionals to use ontology-based informatics tools that allow the execution of this type of logic, and the need for common methods to be developed to distribute computable best practice rules to ensure rapid dissemination of evidence, better translating research into practice.

Study design for prospective studies often requires tradeoffs between tightly specifying the inclusion/exclusion criteria for the study has to be tempered by how each addition to the criteria will influence the rate of recruitment of patients to the study. Once a study has been designed, data-driven recruitment can facilitate rapid recruitment of participants into the trial by identifying individuals we completely match the inclusion and avoid the exclusion criteria for the trial. Once the patients have been recruited, their outcomes can be followed through a prospective and timed recording of data from clinical data repositories, thereby lessening the data entry burden for the trialist. The combination of more rapid recruitment and faster acquisition of outcomes data for the clinical trial will speed the time that it takes to complete clinical trials, thereby bringing evidence more quickly to the bedside.

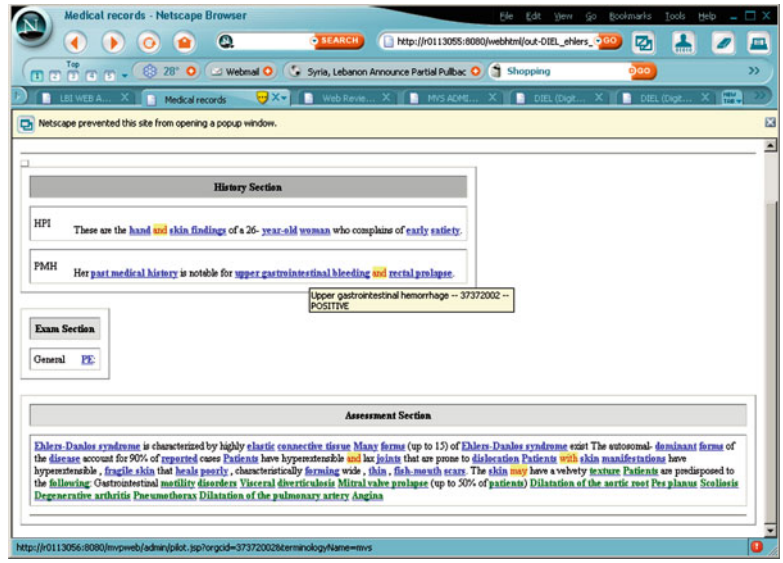


Fig. 9.10 The SNOMED CT mappings of clinical content that can be linked to clinical records for secondary educational use of clinical data at the point of care. The blue concepts are positive assertions, the red are negative assertions, and the green are uncertain assertions

Retrospective studies can often be fully automated. In our example shown below, through linking the NLP-based extract of our clinical records which are encoded using SNOMED CT, RxNorm, and LOINC to our clinical data warehouse that holds all of the structured data (e.g., laboratory results), we can run whole studies and get answers back in minutes rather than years (see Fig. 9.11).

This has the potential to change completely the clinical research paradigm. When clinical trials can be designed and executed in 5 or 10 min, many more questions can be answered, and our understanding of best practice will evolve much quicker harnessing the collective clinical expertise of a greatly expanded set of clinical researchers (i.e., potentially the set of all clinical practitioners). We call these fully automated electronic studies *e*Studies. The interface below we have named the Semantic Biome, and it facilitates research trials that link our structured data from our data warehouse with the ontology-based encodings from our traditionally unstructured data (e.g., Clinical Notes, Discharge Summaries, Radiology Reports, and Pathology Reports).

Data Interoperability

This requires a common data infrastructure based on nationally standard healthcare models for data interchange, linking clinical data from standard ontologies such as SNOMED CT, LOINC, and RxNorm using a common and standard method of logical binding. In the end, one seeks to have a common representation of knowledge across all knowledge types. Facts can be given context by model, patient, document, section (e.g., history), subsection (e.g., HPI), problem, sentence, phrase, compositional expression, and concept level detail. All data should be date and time stamped. These dates can be built into courses of illness, courses of treatment, and courses of hospitalization. Any value-based data held in free text should be able to be extracted and associated with concepts extracted from the text (e.g., total cholesterol measurement has value 150 mg/dl). This knowledge-based common data infrastructure then becomes the basis for the secondary use of clinical data.

MetaMap

MetaMap is a highly configurable program developed by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text [12]. MetaMap uses a knowledge-intensive approach based on symbolic, natural language processing (NLP) and computational linguistic techniques. MetaMap is one of the foundations

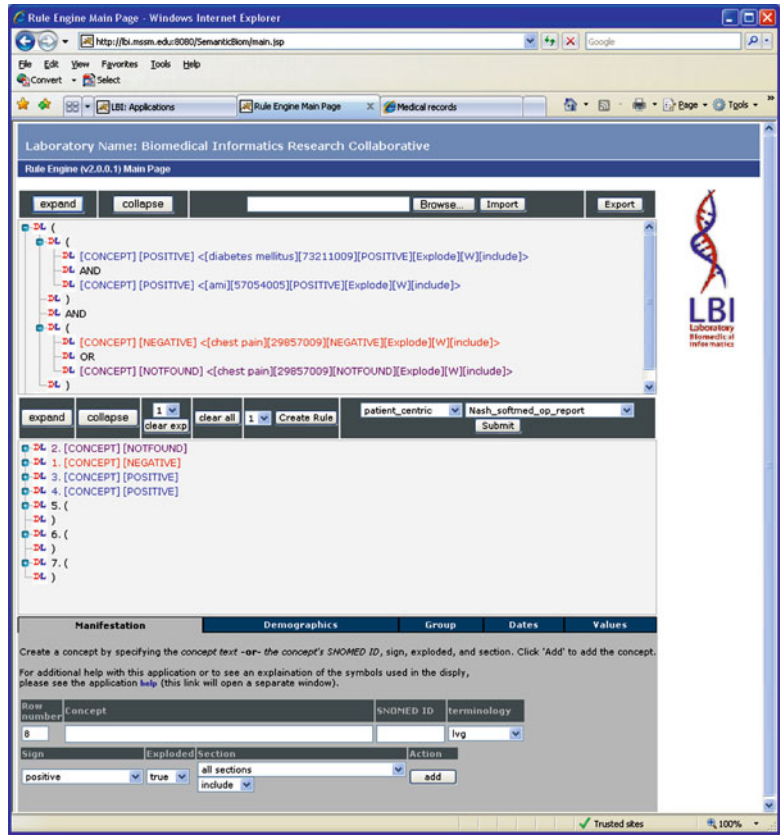


Fig. 9.11 The Semantic Biome – an interface for running electronic studies (*e*Studies). Here we are looking at diabetic patients who had an acute myocardial infarction and did not have chest pain

of NLM's Medical Text Indexer (MTI) which is being applied to both semiautomatic and fully automatic indexing of biomedical literature at NLM. It has also been employed for both information retrieval and data mining applications.

MetaMap [13] is a widely available program providing the concepts identified from the unified medical language system (UMLS) Metathesaurus and associated with the content of biomedical texts. MetaMap arose in the context of an effort to improve information retrieval from biomedical text, specifically the retrieval of MEDLINE/PubMed citations. It provides a link between the text of biomedical literature and the knowledge, including synonymy relationships, embedded in the Metathesaurus. Early MetaMap development was guided by linguistic principles which provided both a rigorous foundation and a flexible architecture in which to explore mapping strategies and their applications. A system diagram showing MetaMap processing is depicted in figure X. Input text undergoes a lexical/syntactic analysis consisting of:

- Tokenization, sentence boundary determination, and acronym/abbreviation identification
- Part-of-speech tagging
- Lexical lookup of input words in the SPECIALIST lexicon
- A final syntactic analysis consisting of a shallow parse in which phrases and their lexical heads are identified by the SPECIALIST minimal commitment parser. Each phrase found by this analysis is further analyzed by the following processes:
- Variant generation, in which variants of all phrase words are determined (usually by a table lookup)
- Candidate identification, in which intermediate results consisting of Metathesaurus strings, called candidates, matching some phrase text are computed and evaluated as to how well they match the input text

- Mapping construction, in which candidates found in the previous step are combined and evaluated to produce a final result that best matches the phrase text
- Optionally, word sense disambiguation (WSD), in which mappings involving concepts that are semantically consistent with surrounding text are favored

The evaluation performed on both the candidates and the final mappings is a linear combination of four linguistically inspired measures: centrality, variation, coverage, and cohesiveness. The evaluation process begins by focusing on the association, or mapping, of input text words to words of the candidates. Centrality, the simplest of the measures, is a Boolean value which is one if the linguistic head of the input text is associated with any of the candidate words. The variation measure is the average of the variation between all text words and their matching candidate words. Coverage and cohesiveness measure how much of the input text is involved in the mapping (the coverage) and in how many chunks of contiguous text (the cohesiveness). The four measures are combined linearly giving coverage and cohesiveness twice the weight of centrality and variation, and the result is normalized to a value between 0 and 1000. MetaMap is highly configurable across dimensions, including:

- Data options, which choose the vocabularies and data model to use.
- Output options, which determine the nature and format of the output generated by MetaMap.
- Processing options, which control the algorithmic computations to be performed by MetaMap. The data options allow the user to choose the UMLS data (e.g., 2009 for the 2009AA release) for use by MetaMap, and the desired level of filtering to employ.

MetaMap's relaxed data model employs:

- Lexical filtering, which excludes most Metathesaurus strings mapped to a concept which are essentially identical to another string for the same concept
- Manual filtering, which excludes unnecessarily ambiguous terms, as determined by a detailed annual study

MetaMap's strict model supplements the above filtering regimen with:

- Syntactic filtering, which excludes complex expressions with underlying grammatical substances, which MetaMap would normally be unable to find anyway because they span multiple phrases

Typical output options include:

- Hiding or displaying the semantic types or concept unique identifiers (CUI) of all displayed concepts
- Hiding or displaying candidates or mappings, where MetaMap will not even compute these elements unless some other option requires them
- Generating XML output rather than the default human-readable output
- Excluding or restricting output to concepts of specified semantic types
- Excluding or restricting output to concepts drawn from specified vocabularies

Some of MetaMap's most useful processing options include:

- Controlling the types of derivational variants used in lexical variant generation (no variants at all, adjective/noun variants only, or all variants)
- Turning on and off MetaMap's WSD module
- Term processing, which causes MetaMap to process each input record, no matter how long, as a single phrase, in order to identify more complex Metathesaurus terms
- Allowing overmatches so that, for example, the input text medicine will map to any concept containing the word medicine, medical, or any other variant of medicine
- Allowing concept gaps so that, for example, the text obstructive apnea will map to concepts "obstructive sleep apnoea" and "obstructive neonatal apnea," which are considered too specific for normal processing

Note that the combination of the last three options (together with hiding the mappings) is known as MetaMap's browse mode. It is generally used to explore the Metathesaurus both broadly and deeply as opposed to the more normal mode in which the "best match" to the input text is sought. Details of all aspects of MetaMap processing can be found in the technical documents at the MetaMap portal.

As mentioned earlier, the final phase of MetaMap's lexical/ syntactic processing involves computing a shallow parse, dividing the input text into phrases, which form the basis of MetaMap's subsequent processing. Each phrase's human-readable output by default consists of three parts:

- The input phrase itself.
- The candidates, a list of intermediate results consisting of Metathesaurus strings matching some or all of the input text. In addition, the preferred name of each candidate is displayed in parentheses if it differs from the candidate, and the semantic type of the candidate is also shown.
- The final mappings, consisting of combinations of candidates matching as much of the input phrase as possible.

Most elements of the human-readable output can be shown or hidden based on how one chooses the MetaMap options. By default, MetaMap displays only those mappings that receive the highest score.

MetaMap possesses a number of strengths and weaknesses. Among its strengths are its thoroughness, characterized by its aggressive generation of word variants, and its linguistically principled approach to its lexical and syntactic analyses as well as its evaluation metric for scoring and ranking concepts. It is also capable of constructing partial, compound mappings when a single concept is insufficient to characterize the input text. MetaMap is highly configurable; its behavior can be easily customized depending on the task to be addressed. Finally, because its lexicon and target vocabulary can be replaced with others from another domain, it has the property of domain independence.

One of MetaMap's weaknesses is that it can be applied only to English text. MetaMap's English-centric nature is evident throughout its implementation, not just in its lexical and syntactic algorithms. Also, a negative consequence of its thoroughness is that it is relatively slow. In its current implementation, it is not appropriate for real-time use, and it would require a major fine-grained parallel reimplementation in order to overcome this weakness. The efficiency enhancements described below in the algorithm tuning section, with rare exceptions, allow MetaMap to process a given MEDLINE citation in well under a minute, although complex phrases, for example:

> from filamentous bacteriophage f1 PCR polymerase-chain reaction PDB Protein Data Bank PSTI human pancreatic secretory trypsin inhibitor RBP retinol-binding protein SPR surface plasmon resonance TrxA E. coli thioredoxin can still require hours of computation because they generate many hundreds of thousands of potential mappings. It is examples such as these that make it clear that reimplementing the MetaMap algorithm to process phrases in parallel would not in general be sufficient to sanction the use of MetaMap for real-time or high-volume applications. That would require subphrasal parallelization, which for mapping construction represents a nontrivial challenge. Although MetaMap was originally designed for tasks that can easily be accomplished using our scheduler, which employs multiple servers to provide document-level parallelization, it is likely that we will undertake a fine-grained parallelization effort in the future.

Perhaps MetaMap's greatest weakness is its reduced accuracy in the presence of ambiguity. MetaMap employs a word sense disambiguation (WSD) algorithm to reduce ambiguity, but it is clear that further disambiguation efforts will be needed to solve the problem satisfactorily, especially as the Metathesaurus is becoming ever more ambiguous.

After some experience with MetaMap, it became clear that the method could be applied to tasks other than retrieval, namely, text mining, classification, question answering, knowledge discovery, and concept-based indexing. In addition, research efforts involving MetaMap have extended to groups outside the National Library of Medicine (NLM). Requests for access to MetaMap from the biomedical informatics community grew over the years.

MetaMap has been used by NLM researchers and outside users since 1994 and is currently available via web access, a downloadable Java implementation (MMTx), an application programming interface, and a downloadable version of the complete Quintus Prolog implementation of MetaMap (see Fig. 9.12). MetaMap was originally developed using Quintus Prolog, which is available from and maintained by the Swedish

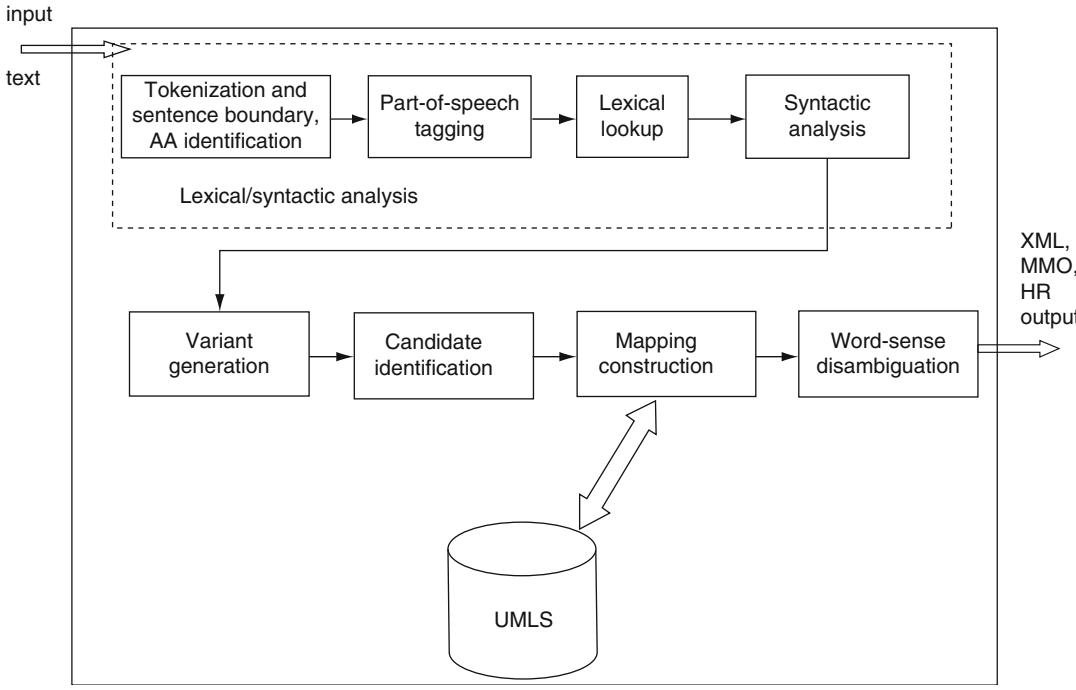


Fig. 9.12 MetaMap system diagram

Institute of Computer Science (http://www.sics.se). MetaMap and MMTx are two versions of the same program. MetaMap, the original program, was developed using Prolog because the language lent itself well to prototyping natural language processing (NLP) applications. The NLM created the Java-based MMTx as a way to distribute MetaMap while separating development from production efforts and because of its platform independence and zero cost. The NLM has since discovered that, due to MMTx's tokenization/lexicalization routines, the two programs produce slightly different results despite concerted efforts to reconcile them. They also learned that almost no MetaMap users modify the code for which they would incur Prolog licensing fees. These factors make it unnecessary to maintain two versions of the program, and as Prolog provides a better development environment, the NLM is phasing out MMTx by freezing its implementation, limiting development to bug fixes.

NEGEX/OpenNLP

Narrative reports in medical records contain a wealth of information that may augment structured data for managing patient information and predicting trends in diseases. Pertinent negatives are evident in text but are not usually indexed in structured databases. The objective of the study reported here was to test a simple algorithm for determining whether a finding or disease mentioned within narrative medical reports is present or absent. Dr. Chapman developed a simple regular expression algorithm called NegEx that implements several phrases indicating negation, filters out sentences containing phrases that falsely appear to be negation phrases, and limits the scope of the negation phrases. They compared NegEx against a baseline algorithm that has a limited set of negation phrases and a simpler notion of scope. In a test of 1235 findings and diseases in 1000 sentences taken from discharge summaries indexed by physicians, NegEx had a specificity of

94.5% (versus 85.3% for the baseline), a positive predictive value of 84.5% (versus 68.4% for the baseline) while maintaining a reasonable sensitivity of 77.8% (versus 88.3% for the baseline). They concluded that with little implementation effort a simple regular expression algorithm for determining whether a finding or disease is absent can identify a large portion of the pertinent negatives from discharge summaries [14].

The NegEx algorithm is a simple one for identifying negatives in textual medical records. It was created by Wendy Chapman, Ph.D., and refined in 2003 with 291 new phrases added to the "negation phrase list." You input a text document to try and see if the NegEx algorithm will find a phrase or phrases within the document.

NegEx is different from the algorithms that they already had out to scan documents. The baseline algorithm that was already out "negates everything from the occurrence of the negation until the end of the sentence," while "NegEx differs between two basic negation types…." NegEx catches things they call double negatives (e.g., "not ruled out") that the original algorithm would miss. It is important for the software to catch these things so that they do not miss anything during their diagnosis and can treat the patient as safe and as quickly as possible. In the medical field, clerical errors contribute to a lot of things that go bad, so not missing these important phrases that the original algorithm would miss is very helpful to the medical community.

Here are some numbers to show you how the NegEx algorithm differs from the original baseline one (see Fig. 9.13). "NegEx had a specificity of 94.5% (versus 85.3% for the baseline), a positive predictive value of 84.5% (versus 68.4% for the baseline) while maintaining a reasonable sensitivity of 77.8% (versus 88.3% for the baseline)."

OpenNLP is an organization dedicated to hold open-source projects related to natural language processing. Its primary role is to encourage and facilitate the collaboration of researchers and developers on such projects.

OpenNLP also hosts a variety of Java-based NLP tools which perform sentence detection, tokenization, pos-tagging, chunking and parsing, named-entity detection, and coreference using the OpenNLP Maxent machine learning package [15].

	Baseline algorithm (%)	NegEx (%)
Sensitivity	88.27	77.84
Specificity	85.27	94.51
PPV	68.42	84.49
NPV	93.01	91.73

Fig. 9.13 The accuracy of the NegEx algorithm

Watson

Watson, named after IBM founder Thomas J. Watson, was built by a team of IBM scientists who set out to accomplish a grand challenge which was to build a computing system that rivals a human's ability to answer questions posed in natural language with speed, accuracy, and confidence (see Fig. 9.14) [16]. Watson competed successfully on Jeopardy against the show's two most successful and celebrated contestants, Ken Jennings and Brad Rutter, on February 14, 15, and 16, of 2011. The Jeopardy format provides the ultimate challenge because the game's clues involve analyzing subtle meaning, irony, riddles, and other language complexities in which humans excel and computers traditionally do not.

Beyond Jeopardy, the technology behind Watson can be adapted to solve problems and drive progress in various fields. The computer has the ability to sift through vast amounts of data and return precise answers, ranking its confidence in its answers. The technology could be applied in areas such as healthcare, to help accurately diagnose patients.

Watson is a significant achievement in the scientific field of Question and Answering, also known as "QA." The Watson software is powered by an IBM POWER7® server optimized to handle the massive number of tasks that Watson must perform at rapid speeds to analyze complex language and deliver correct responses to Jeopardy clues. The system incorporates a number of proprietary technologies for the specialized demands of processing an enormous number of concurrent tasks and data while analyzing information in real time.

IBM's Watson computer competed successfully with two of the best people to have played Jeopardy. This was an impressive display of speech recognition, NLP, data mining, reasoning,

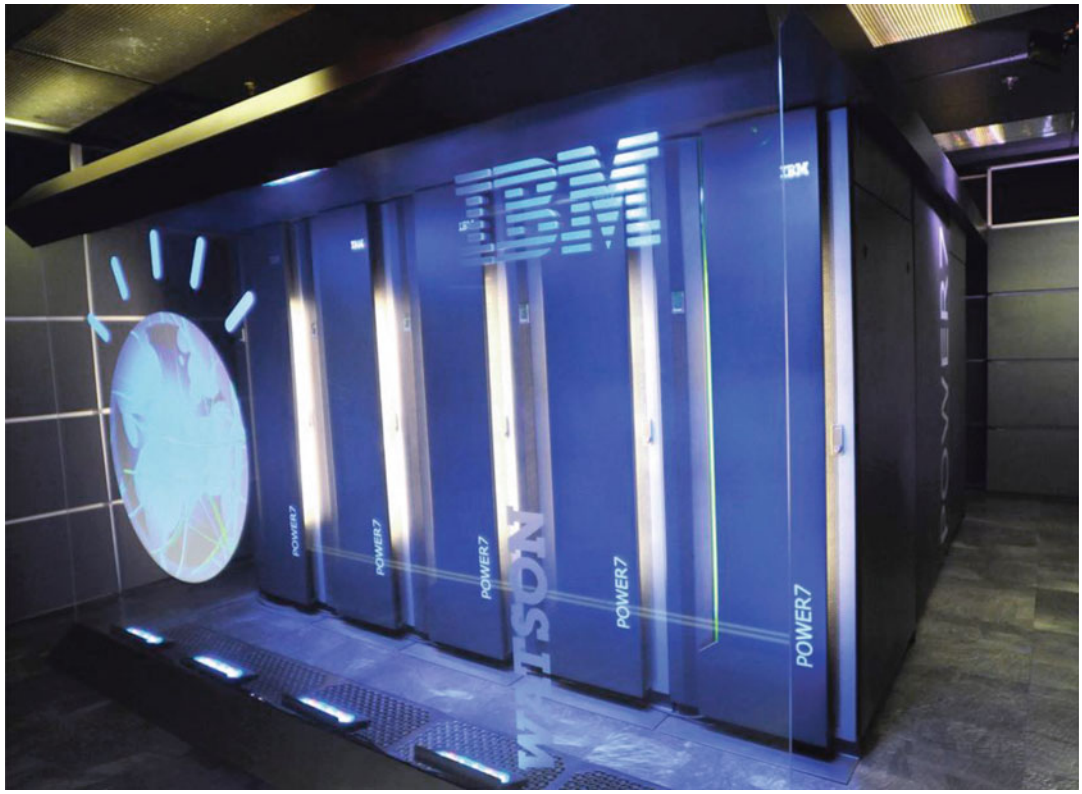


Fig. 9.14 IBM Watson Computer

and of parallel computing. IBM hopes to make an impact on healthcare as one of their next endeavors using the Watson computer.

Watson uses the Apache Unstructured Information Management Application (UIMA) to scale out its natural language processing in parallel across its POWER7 processors, allowing Watson to perform thousands of analytical computations simultaneously across the server cluster to answer each question as fast as possible.

UIMA

Unstructured Information Management Applications are software systems that analyze large volumes of unstructured information in order to discover knowledge that is relevant to an end user [17].

Below is depicted:

- Frameworks
- Components
- Infrastructure,

all available free under the Apache II Open Source license. The dashed-line boxes in Fig. 9.15 which depicts the UIMA environment are placeholders for possible future additions.

UIMA enables applications to be decomposed into components, for example, "language identification" =>"language specific segmentation" =>"sentence boundary detection" =>"entity detection (person/place names, etc.)." Each component implements interfaces defined by the framework and provides self-describing metadata via XML descriptor files. The framework manages these components and the data flow between them. Components are written in Java or C++, and the data that flows between components are designed for efficient mapping between these languages.

UIMA additionally provides capabilities to wrap components into network services and can scale to very large volumes by replicating processing pipelines over a cluster of networked nodes.

Apache UIMA is an Apache-licensed open-source implementation of the UIMA specification (that specification is, in turn, being developed

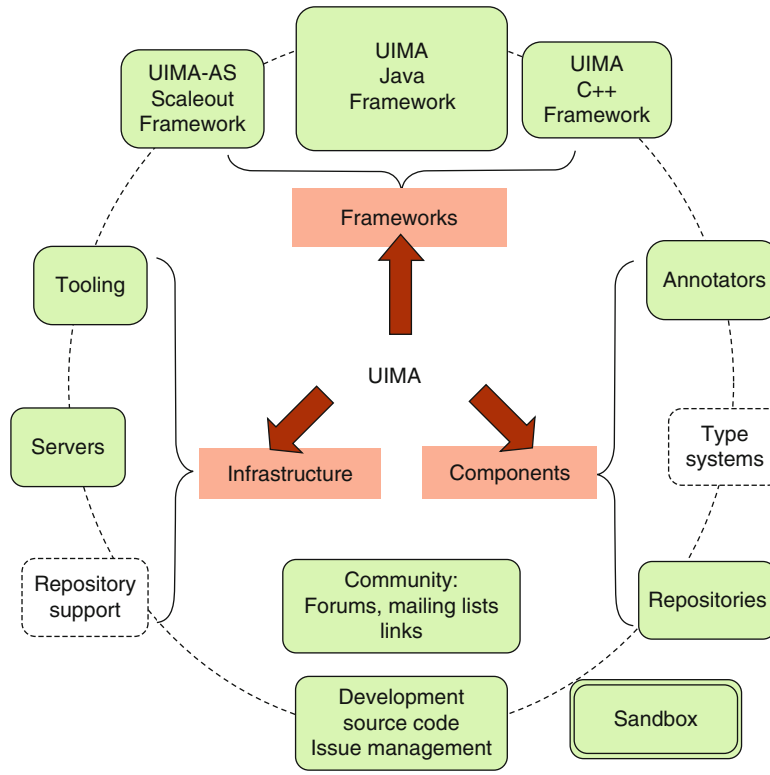


Fig. 9.15 UIMA environment and frameworks, infrastructure, and components. This diagram also depicts relationships to outside communities, development practices, and a test sandbox

concurrently by a technical committee within OASIS, a standards organization). Apache invites and encourages you to participate in both the implementation and specification efforts.

UIMA Introduction

Unstructured information represents the largest, most current, and fastest growing source of information available to businesses and governments. The web is just the tip of the iceberg.

Consider the mounds of information hosted around the world and across different media including text, voice, and video. The high-value content in these vast collections of unstructured information is, unfortunately, buried in a lot of noise. Searching for what you need or doing sophisticated data mining over unstructured information sources presents new challenges.

An Unstructured Information Management (UIM) Application may be generally characterized as a software system that analyzes large volumes of unstructured information (text, audio, video, images, etc.) to discover, organize, and deliver relevant knowledge to the client or application end user (see Fig. 9.16).

An example would be an application that processes millions of medical abstracts to discover critical drug interactions. Another example would be an application that processes tens of millions of documents to discover key evidence indicating an emergent infectious disease.

First and foremost, the unstructured data must be analyzed to interpret, detect, and locate concepts of interest, for example, named entities like persons, organizations, locations, facilities, products, etc., that are not explicitly tagged or annotated in the original artifact. More challenging analytics may detect things like opinions, complaints, findings, disorders, or facts. The list of concepts important for applications to discover in unstructured content is large, varied, and often domain specific.

Many different component analytics may solve different parts of the overall analysis task. These component analytics must interoperate and must be easily combined to facilitate the development of UIM applications.

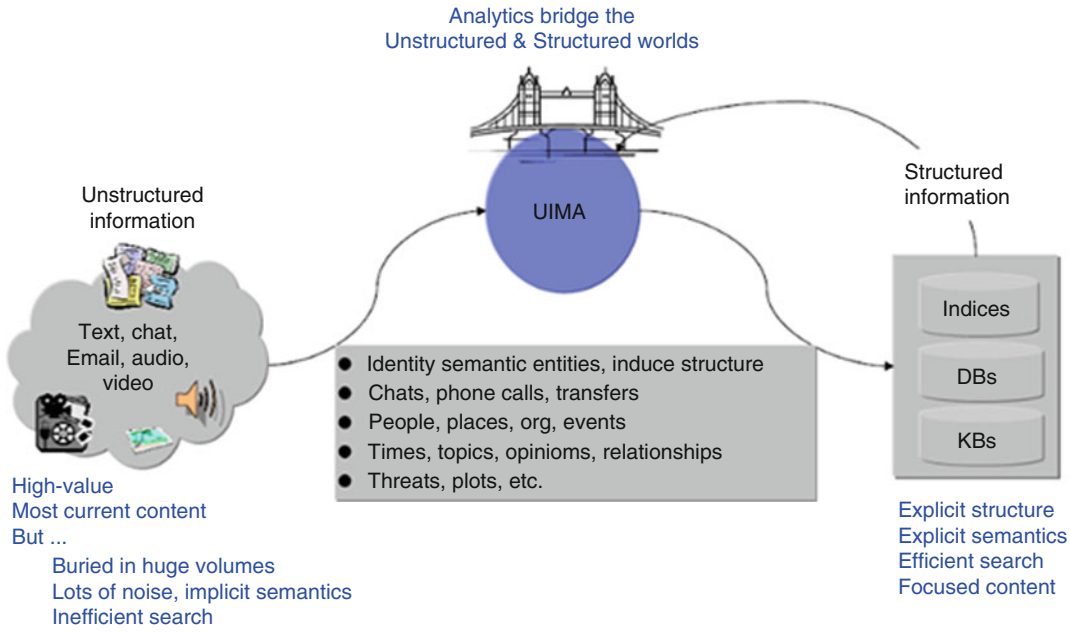


Fig. 9.16 UIMA helps you build the bridge between the unstructured and structured worlds

The result of analysis is used to populate structured forms so that conventional data processing and search technologies like search engines, database engines, or OLAP (online analytical processing, or data mining) engines can efficiently deliver the newly discovered content in response to the client requests or queries.

In analyzing unstructured content, UIM applications make use of a variety of analysis technologies including:

- Statistical and rule-based natural language processing (NLP)
- Information retrieval (IR)
- Machine learning
- Ontologies
- Automated reasoning
- Knowledge sources (e.g., CYC, WordNet, FrameNet, etc.)

Specific analysis capabilities using these technologies are developed independently using different techniques, interfaces, and platforms. The bridge from the unstructured world to the structured world is built through the composition and deployment of these analysis capabilities. This integration is often a costly challenge.

The Unstructured Information Management Architecture (UIMA) is an architecture and software framework that helps you build that bridge. It supports creating, discovering, composing, and deploying a broad range of analysis capabilities and linking them to structured information services.

UIMA allows development teams to match the right skills with the right parts of a solution and helps enable rapid integration across technologies and platforms using a variety of different deployment options. These range from tightly coupled deployments for high-performance, single-machine, embedded solutions to parallel and fully distributed deployments for highly flexible and scalable solutions.

UIMA is a software architecture which specifies component interfaces, data representations, design patterns, and development roles for creating, describing, discovering, composing, and deploying multimodal analysis capabilities.

The UIMA framework provides a run-time environment in which developers can plug in their UIMA component implementations and with which they can build and deploy UIM applications.

Terminology Servers and Services

DTS

The Apelon DTS (Distributed Terminology System) is an integrated set of components that provides comprehensive terminology services in distributed application environments [18]. Currently in use in many leading healthcare organizations, the DTS provides support for national (and international) data standards as well as local vocabularies, necessary foundations for comparable and interoperable health information. Typical applications for DTS include clinical data entry, results review, problem-list and code-set management, guideline creation, decision support, and information retrieval.

Key DTS features include:

- *High performance*: Concurrent access to multiple, interconnected terminologies
- *Comprehensive*: Extensive terminology knowledge base with a unified, consistent object model
- *Data normalization*: Matching of text input to standardized terms and concepts via word order analysis, word stemming, spelling correction, and term completion
- *Code translation*: Mapping of clinical data to standard coding systems such as ICD-9 and CPT®
- *Class queries*: Hierarchy interrogation for decision support and outcomes analysis
- *Semantic navigation*: Browsing of a rich set of hierarchical and nonhierarchical relationships between concepts for improved quality in data entry and information retrieval
- *Semantic classification*: Creation, management, and comparison of concept extensions which are consistent with formal semantic models such as that used in SNOMED CT®
- *Subsetting*: Creation of individualized subsets of terminologies using advanced Boolean logic techniques
- *Workflow*: Management and tracking of modeling efforts in large, distributed projects
- *Localization*: Addition of local concepts, synonyms, codes, properties, and interconcept associations to connect local content to standard terminologies

DTS provides APIs and management applications for both Java and Microsoft .NET environments. The extensible DTS Editor enables the enhancement of the DTS Knowledge Base by adding new content and localizing it for specific business, professional, or cultural needs, such as noting that "Black Creek disease" is a synonym for amebic dysentery. The DTS Browser permits easy access and review of terminologies from any Internet browser.

DTS in now available as an open-source project on Source Forge.

Apelon's TermWorks is an innovative data mapping solution which brings powerful terminology capabilities directly to the desktop. TermWorks combines Microsoft® Excel® spreadsheet software with web services-based terminology processing to give organizations comprehensive mapping capability without the high cost of hardware and software acquisition, installation, integration, maintenance, and support.

Apelon's TermWorks Excel plug-in extends this familiar application with advanced terminology capabilities such as search and concept navigation, operating on industry-standard terminologies such as SNOMED CT®, CPT®, and ICD-9-CM.

TermWorks uses web services to access remote servers hosted by Apelon. These high-performance systems maintain the most recent version of all standard healthcare terminologies and provide the processing power behind TermWorks' sophisticated matching and searching algorithms.

Organizations can also use TermWorks web services directly to include advanced terminology capabilities into their applications. These applications can take advantage of rich vocabulary features and industry-standard terminologies.

iNLP Server

The iNLP server is written in the .NET framework and is exposed via a set of web services with an easy to learn and use Web Services Description Language (WSDL) interface shown below:

```
<?xml version="1.0" encoding="UTF-8"?>
<wsdl:definitions xmlns:wsdl="http://
schemas.xmlsoap.org/wsdl/"
xmlns:ns1="http://org.apache.axis2/
xsd" xmlns:ns="http://ws.mrc.cbi.LBI.
edu" xmlns:wsaw="http://www.w3.
org/2006/05/addressing/wsdl"
xmlns:http="http://schemas.xmlsoap.
org/wsdl/http/" xmlns:xs="http://www.
w3.org/2001/XMLSchema"
xmlns:soap="http://schemas.xmlsoap.
org/wsdl/soap/" xmlns:mime="http://
schemas.xmlsoap.org/wsdl/mime/"
xmlns:soap12="http://schemas.xmlsoap.
org/wsdl/soap12/"
targetNamespace="http://ws.mrc.cbi.
LBI.edu">
<wsdl:documentation>
Please Type your service description
here
</wsdl:documentation>
  <wsdl:types>
    <xs:schema
attributeFormDefault="qualified"
elementFormDefault="qualified"
targetNamespace="http://ws.mrc.cbi.
mssm.edu">
      <xs:element
name="getCodifiedMedicalRecord">
        <xs:complexType>
          <xs:sequence>
            <xs:element minOccurs="0"
name="runId" type="xs:int"/>
            <xs:element minOccurs="0"
name="appId" type="xs:int"/>
            <xs:element minOccurs="0"
name="patientId" type="xs:int"/>
            <xs:element minOccurs="0"
name="terminology" nillable="true"
type="xs:string"/>
            <xs:element minOccurs="0"
name="plainTextRecord" nillable="true"
type="xs:string"/>
            <xs:element minOccurs="0"
name="docName" nillable="true"
type="xs:string"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="getCodifiedMedic
alRecordResponse">
        <xs:complexType>
          <xs:sequence>
            <xs:element minOccurs="0"
name="return" nillable="true"
type="xs:string"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="getCodifiedMedic
alRecordWithSMSAlert">
        <xs:complexType>
          <xs:sequence>
            <xs:element minOccurs="0"
name="runId" type="xs:int"/>
            <xs:element minOccurs="0"
name="appId" type="xs:int"/>
            <xs:element minOccurs="0"
name="patientId" type="xs:int"/>
            <xs:element minOccurs="0"
name="terminology" nillable="true"
type="xs:string"/>
            <xs:element minOccurs="0"
name="plainTextRecord" nillable="true"
type="xs:string"/>
            <xs:element minOccurs="0"
name="docName" nillable="true"
type="xs:string"/>
            <xs:element minOccurs="0"
name="toNumber" nillable="true"
type="xs:string"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="getCodifiedMedic
alRecordWithSMSAlertResponse">
        <xs:complexType>
          <xs:sequence>
            <xs:element minOccurs="0"
name="return" nillable="true"
type="xs:string"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="getCodifiedMedic
alRecordWithSMSAlertWithName">
```

```
      <xs:complexType>
        <xs:sequence>
          <xs:element minOccurs="0" name="runId" type="xs:int"/>
          <xs:element minOccurs="0" name="appId" type="xs:int"/>
          <xs:element minOccurs="0" name="patientId" type="xs:int"/>
          <xs:element minOccurs="0" name="terminology" nillable="true"
type="xs:string"/>
          <xs:element minOccurs="0" name="plainTextRecord"nillable="true"
type="xs:string"/>
          <xs:element minOccurs="0" name="docName" nillable="true"
type="xs:string"/>
          <xs:element minOccurs="0" name="toNumber" nillable="true"
type="xs:string"/>
          <xs:element minOccurs="0" name="firstName" nillable="true"
type="xs:string"/>
          <xs:element minOccurs="0" name="lastName" nillable="true"
type="xs:string"/>
        </xs:sequence>
      </xs:complexType>
    </xs:element>
    <xs:element name="getCodifiedMedicalRecordWithSMSAlertWithNameResponse">
      <xs:complexType>
        <xs:sequence>
          <xs:element minOccurs="0" name="return"nillable="true"
type="xs:string"/>
        </xs:sequence>
      </xs:complexType>
    </xs:element>
  </xs:schema>
 </wsdl:types>
 <wsdl:message name="getCodifiedMedicalRecordWithSMSAlertRequest">
  <wsdl:part name="parameters" element="ns:getCodifiedMedicalRecordWithSMSAler
t"/>
 </wsdl:message>
 <wsdl:message name="getCodifiedMedicalRecordWithSMSAlertResponse">
 <wsdl:part name="parameters" element="ns:getCodifiedMedicalRecordWithSMSAlertR
esponse"/>
 </wsdl:message>
 <wsdl:message name="getCodifiedMedicalRecordWithSMSAlertWithNameRequest">
  <wsdl:part name="parameters" element="ns:getCodifiedMedicalRecordWithSMSAler
tWithName"/>
 </wsdl:message>
 <wsdl:message name="getCodifiedMedicalRecordWithSMSAlertWithNameResponse">
  <wsdl:part name="parameters" element="ns:getCodifiedMedicalRecordWithSMSAler
tWithNameResponse"/>
 </wsdl:message>
 <wsdl:message name="getCodifiedMedicalRecordRequest">
  <wsdl:part name="parameters" element="ns:getCodifiedMedicalRecord"/>
 </wsdl:message>
```

```
  <wsdl:message name="getCodifiedMedicalRecordResponse">
    <wsdl:part name="parameters" element="ns:getCodifiedMedicalRecordResponse"/>
  </wsdl:message>
  <wsdl:portType name="MRCodifierPortType">
    <wsdl:operation name="getCodifiedMedicalRecordWithSMSAlert">
      <wsdl:input message="ns:getCodifiedMedicalRecordWithSMSAlertRequest" wsaw:Acti
on="urn:getCodifiedMedicalRecordWithSMSAlert"/>
      <wsdl:output message="ns:getCodifiedMedicalRecordWithSMSAlertResponse" wsa
w:Action="urn:getCodifiedMedicalRecordWithSMSAlertResponse"/>
    </wsdl:operation>
    <wsdl:operation name="getCodifiedMedicalRecordWithSMSAlertWithName">
    <wsdl:input message="ns:getCodifiedMedicalRecordWithSMSAlertWithNameRequest"
wsaw:Action="urn:getCodifiedMedicalRecordWithSMSAlertWithName"/>
      <wsdl:output message="ns:getCodifiedMedicalRecordWithSMSAlertWithNameRespo
nse" wsaw:Action="urn:getCodifiedMedicalRecordWithSMSAlertWithNameResponse"/>
    </wsdl:operation>
    <wsdl:operation name="getCodifiedMedicalRecord">
      <wsdl:input message="ns:getCodifiedMedicalRecordRequest" wsaw:Action="urn:
getCodifiedMedicalRecord"/>
      <wsdl:output message="ns:getCodifiedMedicalRecordResponse" wsaw:Action="ur
n:getCodifiedMedicalRecordResponse"/>
    </wsdl:operation>
  </wsdl:portType>
  <wsdl:binding name="MRCodifierSoap11Binding" type="ns:MRCodifierPortType">
    <soap:binding transport="http://schemas.xmlsoap.org/soap/http"
style="document"/>
    <wsdl:operation name="getCodifiedMedicalRecordWithSMSAlertWithName">
      <soap:operation soapAction="urn:getCodifiedMedicalRecordWithSMSAlertWithNa
me" style="document"/>
      <wsdl:input>
        <soap:body use="literal"/>
      </wsdl:input>
      <wsdl:output>
        <soap:body use="literal"/>
      </wsdl:output>
    </wsdl:operation>
    <wsdl:operation name="getCodifiedMedicalRecordWithSMSAlert">
      <soap:operation soapAction="urn:getCodifiedMedicalRecordWithSMSAlert"
style="document"/>
      <wsdl:input>
        <soap:body use="literal"/>
      </wsdl:input>
      <wsdl:output>
        <soap:body use="literal"/>
      </wsdl:output>
    </wsdl:operation>
    <wsdl:operation name="getCodifiedMedicalRecord">
      <soap:operation soapAction="urn:getCodifiedMedicalRecord"
style="document"/>
```

```
      <wsdl:input>
        <soap:body use="literal"/>
      </wsdl:input>
      <wsdl:output>
        <soap:body use="literal"/>
      </wsdl:output>
    </wsdl:operation>
  </wsdl:binding>
  <wsdl:binding name="MRCodifierSoap12Binding" type="ns:MRCodifierPortType">
    <soap12:binding transport="http://schemas.xmlsoap.org/soap/http"
style="document"/>
    <wsdl:operation name="getCodifiedMedicalRecordWithSMSAlertWithName">
      <soap12:operation soapAction="urn:getCodifiedMedicalRecordWithSMSAlertWith
Name" style="document"/>
      <wsdl:input>
        <soap12:body use="literal"/>
      </wsdl:input>
      <wsdl:output>
        <soap12:body use="literal"/>
      </wsdl:output>
    </wsdl:operation>
    <wsdl:operation name="getCodifiedMedicalRecordWithSMSAlert">
      <soap12:operation soapAction="urn:getCodifiedMedicalRecordWithSMSAlert"
style="document"/>
      <wsdl:input>
        <soap12:body use="literal"/>
      </wsdl:input>
      <wsdl:output>
        <soap12:body use="literal"/>
      </wsdl:output>
    </wsdl:operation>
    <wsdl:operation name="getCodifiedMedicalRecord">
      <soap12:operation soapAction="urn:getCodifiedMedicalRecord"
style="document"/>
      <wsdl:input>
        <soap12:body use="literal"/>
      </wsdl:input>
      <wsdl:output>
        <soap12:body use="literal"/>
      </wsdl:output>
    </wsdl:operation>
  </wsdl:binding>
  <wsdl:binding name="MRCodifierHttpBinding" type="ns:MRCodifierPortType">
    <http:binding verb="POST"/>
    <wsdl:operation name="getCodifiedMedicalRecordWithSMSAlertWithName">
      <http:operation location="MRCodifier/getCodifiedMedicalRecordWithSMSAlertWi
thName"/>
      <wsdl:input>
```

```
      <mime:content type="text/xml" part="getCodifiedMedicalRecordWithSMSAlertW
ithName"/>
     </wsdl:input>
     <wsdl:output>
      <mime:content type="text/xml" part="getCodifiedMedicalRecordWithSMSAlertW
ithName"/>
     </wsdl:output>
   </wsdl:operation>
   <wsdl:operation name="getCodifiedMedicalRecordWithSMSAlert">
     <http:operation location="MRCodifier/getCodifiedMedicalRecordWithSMSAler
t"/>
     <wsdl:input>
      <mime:content type="text/xml" part="getCodifiedMedicalRecordWithSMSAler
t"/>
     </wsdl:input>
     <wsdl:output>
      <mime:content type="text/xml" part="getCodifiedMedicalRecordWithSMSAler
t"/>
     </wsdl:output>
   </wsdl:operation>
   <wsdl:operation name="getCodifiedMedicalRecord">
     <http:operation location="MRCodifier/getCodifiedMedicalRecord"/>
     <wsdl:input>
      <mime:content type="text/xml" part="getCodifiedMedicalRecord"/>
     </wsdl:input>
     <wsdl:output>
      <mime:content type="text/xml" part="getCodifiedMedicalRecord"/>
     </wsdl:output>
   </wsdl:operation>
  </wsdl:binding>
  <wsdl:service name="MRCodifier">
   <wsdl:port name="MRCodifierHttpSoap11Endpoint" binding="ns:MRCodifierSoap11Bi
nding">
     <soap:address location="http://.............:8080/axis2/services/
MRCodifier.MRCodifierHttpSoap11Endpoint/"/>
   </wsdl:port>
   <wsdl:port name="MRCodifierHttpSoap12Endpoint" binding="ns:MRCodifierSoap12Bi
nding">
     <soap12:address location="http://.............:8080/axis2/services/
MRCodifier.MRCodifierHttpSoap12Endpoint/"/>
   </wsdl:port>
   <wsdl:port name="MRCodifierHttpEndpoint" binding="ns:MRCodifierHttpBinding">
     <http:address location="http://.............:8080/axis2/services/
MRCodifier.MRCodifierHttpEndpoint/"/>
   </wsdl:port>
  </wsdl:service>
</wsdl:definitions>
```

1. Now we will generate the client for the newly created service by referring the wsdl generated by the Axis2 Server. Open File ->New ->Other… ->Web Services ->Web ServiceClient (see Fig. 9.17).
2. Paste the URL that was copied earlier into the service definition field (see Fig. 9.18).
3. *Next Web Services Interface Option*: click Server:Tomcat v5.5 Server from #2 and make sure web service run time is Apache Axis2 (see Fig. 9.19).
4. *Next Web Services Screen*: Click on the **Client project** hyperlink and enter **MRCodifierWSClient** as the name of the client project. Click OK (see Fig. 9.20).
5. Next page is the Web Services Client Configuration Page: Accept the defaults and click Finish (see Fig. 9.21).
6. Now we are going to write the Java main program to invoke the client stub. Import the MRCodifierClient.java file in the src folder of **MRCodifierWSClient** (see Fig. 9.22).

The iNLP Services provide full text indexing of clinical records or other documentation and return data structures suitable for storing in XML or in a relational database. The services are terminology and language independent. Languages and terminologies are specified in the WSDL, and the appropriate codes and structures are returned.

CTS II

The Common Terminology Services (CTS) II specification was developed as an alternative to a common data infrastructure. Instead of specifying what an external terminology must look like, HL7 has chosen to identify the common functional characteristics that an external terminology must be able to provide. As an example, an HL7 compliant terminology service will need to be able to determine whether a given concept code is valid within the particular resource. Instead of describing a table keyed by the resource identifier and concept code, the CTS specification describes an application programming interface (API) call that takes a resource identifier and concept code as input and returns a true/false value. Each terminology developer is free to implement this API call in whatever way is most appropriate for them.

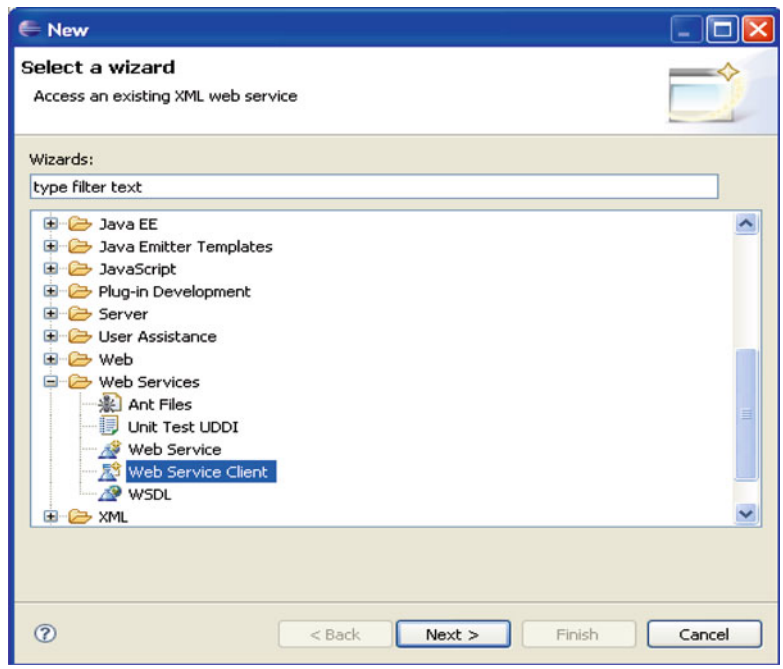


Fig. 9.17 Interface for the newly created service by referring the WSDL generated by the Axis2 server. The user simply follows the commands: Open File ->New ->Other… ->Web Services ->Web ServiceClient

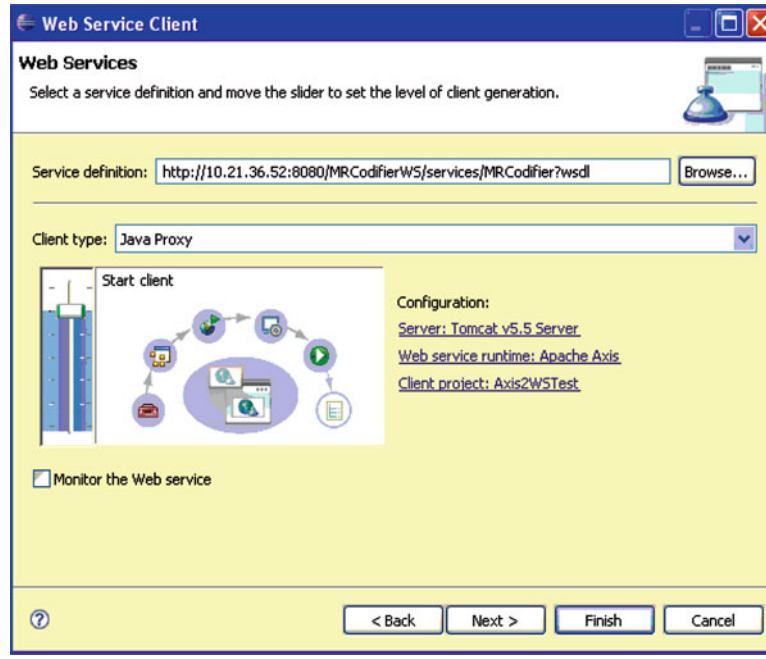


Fig. 9.18 Next Screen in the Web Services Interface: Simply paste the URL that was copied earlier into the service definition field

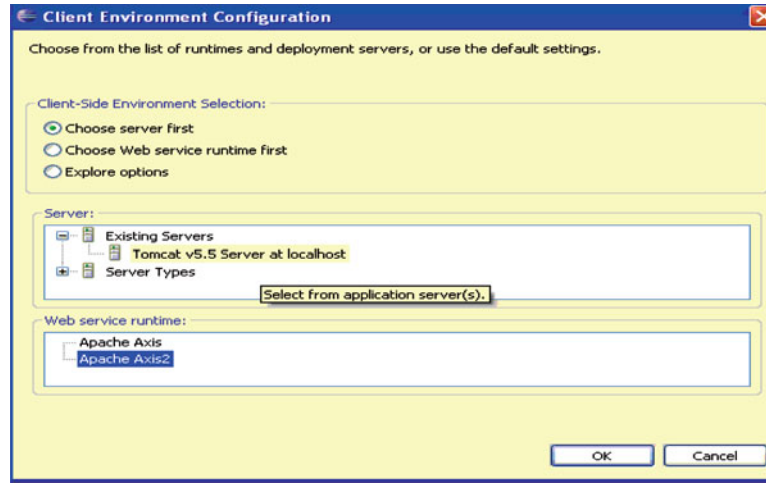


Fig. 9.19 Next Web Services Interface Option: click Server:Tomcat v5.5 Server from #2 and make sure web service run time is Apache Axis2

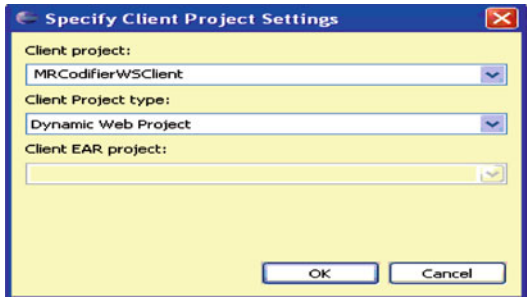


Fig. 9.20 Next Web Services Screen: Click on the **Client project** hyperlink and enter **MRCodifierWSClient** as the name of the client project. Click OK

The CTS specification is not designed to perform the following services:

- The current version of CTS is not intended to be a complete terminology service. The scope of CTS is restricted to the functionality needed to design, implement, and deploy an HL7 Version 3 compliant software package.
- CTS is not intended to be a general purpose query language. It is intended to specify only the specific services needed in the HL7 implementation.

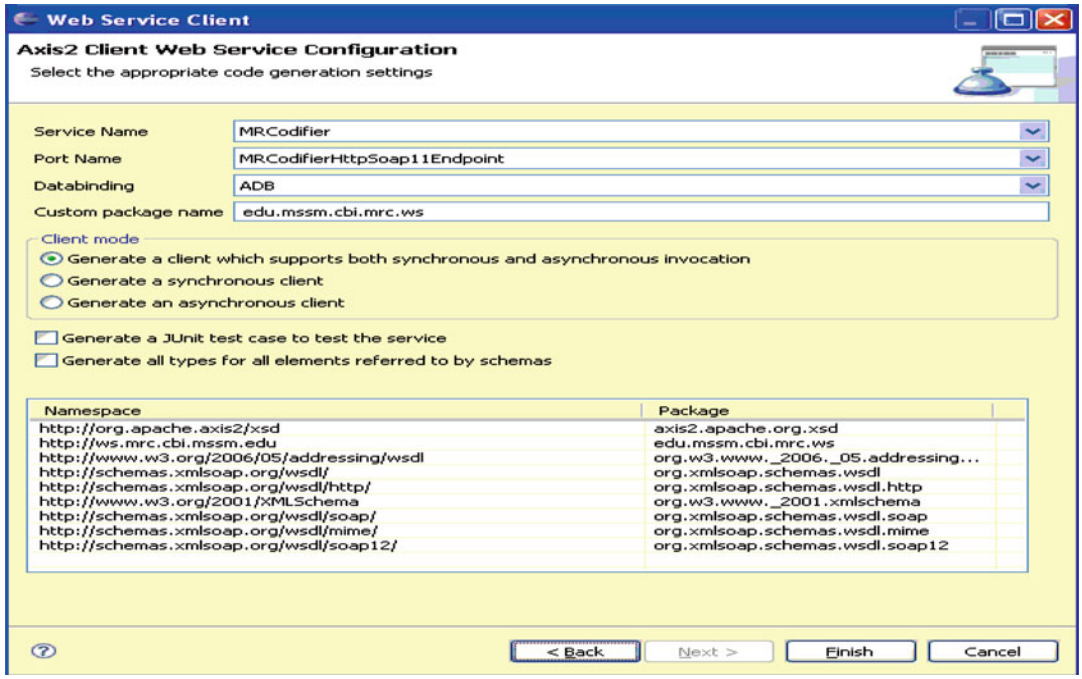


Fig. 9.21 Next page is the Web Services Client Configuration Page. Accept the defaults and click Finish

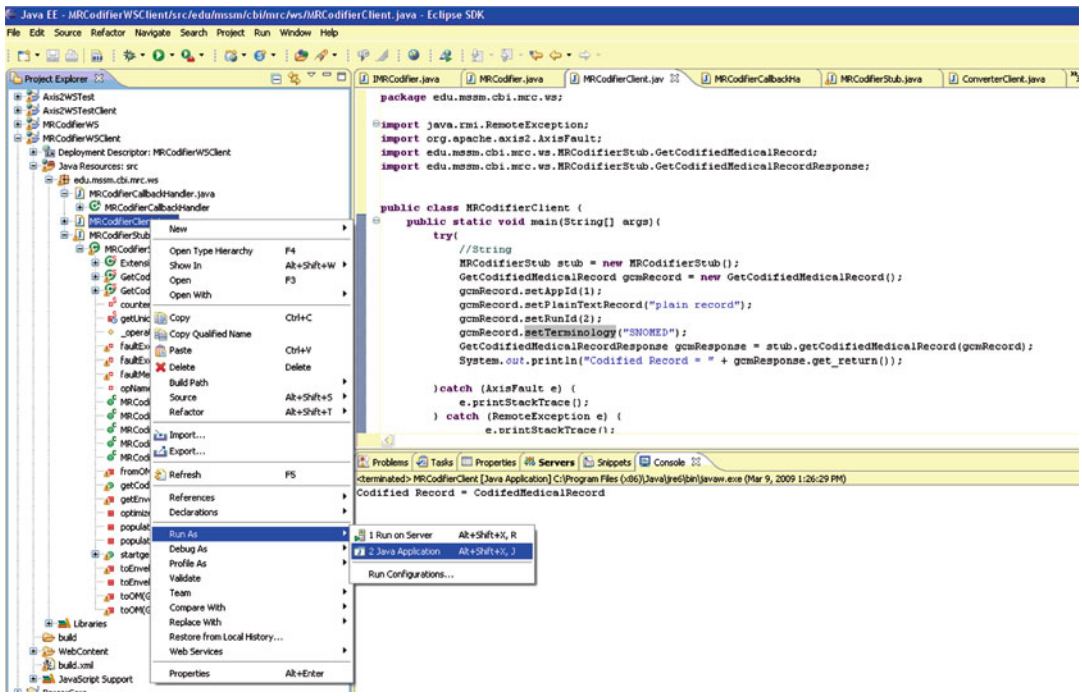


Fig. 9.22 Java main program to invoke the client stub. Import the MRCodifierClient.java file in the src folder of **MRCodifierWSClient**

- CTS II does not specify how the service is to be implemented. It is intentionally silent when it comes to service advertising and discovery, establishing and maintaining connections, and the delivery and routing of messages. It is assumed that a CTS implementation will use the underlying architecture that is most appropriate for the given implementation circumstances.

Secondary Use of Healthcare Data

Clinical data from electronic health records have traditionally contained a small proportion of fixed field data (often obtained from pick lists) and larger quantities of free text. Some EHRs only store images of handwritten or typed notes (e.g., faxed in data). These practices have made it difficult to extract and use electronic health record data for secondary purposes.

These purposes can be categorized into assistance with the practice of medicine, research, and education. The practice of medicine can employ EHR knowledge at the point of care in the form of alerts and expert advice for the clinician, the patient, or their family (caregivers). Ideally these systems could learn from the outcomes associated with the population of patients cared for by a given provider. Research stands to gain substantially by employing EHR data for secondary uses. These can and will range from more intelligent study design where the impact on recruitment can be tested as we add additional criteria to either the inclusion or exclusion criteria for the study. This concept includes data-driven recruitment that will assure that a much higher percentage of participants screened for recruitment to a clinical trial will be found to be appropriate for that trial. For retrospective trials, this technology is able to run fully automated studies and complete trials in minutes rather than years. For prospective studies, this technology is able to track a much broader set of clinical outcomes making more fruitful our research dollar spent. For education, real-time learning systems will be updated with the results of clinical practice, and based on best outcomes this technology is able to educate all physicians in a practice area with information learned from anyone's practice. This continuous learning environment will advance the quality of practice available to all patients.

Examples of successful projects toward the secondary use of healthcare data include projects such as Opticode which is an object-oriented expert system that determines the Evaluation and Management (E&M) billing code based on the contents of a clinician's note in the ambulatory setting. This has been used extensively and has been shown to be as accurate as a physician and in some cases more accurate in its assignments of E&M codes. The Marker Discovery project is a project to identify new markers (genes or proteins) for a disease [19]. This project has identified new molecules that predict and can be used as targets for the identification of patients with a genetically related disorder. The digital image education library (DIEL) is a case-based teaching system where the content is indexed by SNOMED CT for specific retrieval and to enable linkages from and to an EHR [20]. The FDA drug labels project has used the text of the drug labels coded in the HL7 SPL format to provide a set of semantic triples regarding drugs. These triples specify axioms like drug x HasAdverseRxn y with HasFrequency z in HasPopulation a. This information can be used to drive clinical decision support to improve care and avoid adverse drug events [21].

Clinical decision support requires data to drive the decision rules. These data need to be codified in order to trigger the decision rules when appropriate. The inability to harvest the data from the clinical record including the EHR to fuel clinical decision support rules has been named the "curly braces" problem. This stems from the clinical rules containing statements such as:

If the Patient is taking {Diuretic} and their {Serum Potassium} is less than 3.0 mg/dl then hold the {Diuretic}

which often depict curly braces around variables that are needed to feed the rule's trigger model. In this case, the rule assumes that you can tell (a) whether the patient is taking any diuretic and (b) what is the value of the patient's serum potassium? In order to find these answers, the codes in the clinical record must match those in

the clinical repositories that need to serve up the data to trigger this rule. Sound terminological systems that support this type of secondary use of clinical data are necessary to solve the curly braces problem.

Many diagnostic clinical decision support systems use standardized terminology as input to their expert algorithms. Some examples of these are DXplain built by G. Octo Barnett, M.D., and distributed by Massachusetts General Hospital; QMR built by Randy Miller, M.D., and Jack Myers, M.D.; Iliad built at LDS Hospital by Homer R. Warner, MD, and others.

In order for this dream to become a reality, it requires a common data infrastructure into which all clinical data are represented. This requires defining the formalism, and then it requires a method for encoding the clinical data recorded during the normal clinical care workflow into this common representation schema. The formalism to be usable must represent the data at the same level of granularity as is recorded in routine clinical practice.

Conclusion

Terminological systems are where the rubber meets the road. These workhorses of the terminology world process the text, index that same output, create the context and structure around the content that has been discovered, and then serve up the content in support of the secondary use of healthcare data. This is the best and only way to rapidly improve the quality and safety of clinical practice while reducing costs. In so doing we provide better healthcare value for the people and their families who have put their trust in us to provide for them the very best healthcare possible.

Questions

1. Terminological systems include:
   (a) MetaMap
   (b) IHTSDO Workbench
   (c) Negex
   (d) All of the above
2. The secondary use of clinical data:
   (a) Includes all uses of clinical data
   (b) Does not include clinical uses of healthcare data
   (c) Includes natural language processing software
   (d) Does not include decision support
3. Which correctly lists the order from oldest to newest of these terminology development platforms?
   (a) Apelon TDE, IHTSDO Workbench, Protégé
   (b) Protégé, Apelon TDE, IHTSDO Workbench
   (c) Apelon TDE, Protégé, IHTSDO Workbench
   (d) IHTSDO Workbench, Apelon TDE, Protégé
4. Which natural language processor was built by the National Library of Medicine?
   (a) MetaMap
   (b) iNLP
   (c) NegEx
   (d) OpenNLP
5. IBM's Watson Computer employs which of the following?
   (a) SNOMED CT
   (b) HL7 RIM
   (c) UIMA
   (d) ISO TS17117
6. Which of the following is not available through an Open Source license?
   (a) UIMA
   (b) MedLEE
   (c) Protégé
   (d) IHTSDO Workbench
7. Which of the following system(s) employ a description logic classifier?
   (a) Protégé
   (b) Apelon's TDE
   (c) IHTSDO Workbench
   (d) All of the above
   (e) Only a and c
8. The "curly braces" problem in the field of CDS is:
   (a) A syntax problem with formatting rule sets
   (b) A way to call out portions of the rule for which data are missing
   (c) Variables from clinical content that are not easily obtained to trigger rules

(d) Parts of a rule set that hold codified data

9. Terminological systems are important for:
   (a) The development of ontologies
   (b) The use of ontologies for data mining
   (c) The secondary use of clinical data
   (d) All of the above
10. The usability of terminological systems is:
   (a) Important for the system to be usable by knowledge workers
   (b) Important for the system to be usable by subject matter experts
   (c) Important for the accuracy of terminological systems
   (d) Only a and b
   (e) a, b, and c

References

1. http://www.apelon.com/Products/TDE/tabid/100/Default.aspx.
2. http://protege.stanford.edu/.
3. Musen MA, Gennari JH, Eriksson H, Tu SW, Puerta AR. PROTEGE-II: computer support for development of intelligent systems from libraries of components. Medinfo. 1995;8(Pt 1):766–70.
4. http://www.w3.org/TR/owl-guide/.
5. http://www.ihtsdo.org/index.php?id=803.
6. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000:270–4.
7. Joint Commission on Accreditation of Healthcare Organizations. 2005–2006 Standards for ambulatory care. Oakbrook Terrace: JCAHO; 2005.
8. Weed L. The problem-oriented record-its organizing principles and its structure. League Exch. 1975;103:3–6.
9. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, Speroff T. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc. 2006;81(6):741–8.
10. Brown SH, Speroff T, Fielstein EM, Bauer BA, Wahner-Roedler DL, Greevy R, Elkin PL. eQuality: Automatic assessment from narrative clinical reports. Mayo Clin Proc. 2006;81(11):1472–81.
11. Brown SH, Elkin PL, Fielstein E, Speroff T. eQuality for all – extending automated quality measurement from free text. AMIA Annu Symp Proc. 2008;6:71–5.
12. http://mmtx.nlm.nih.gov/.
13. Aronson AR, Lang FM. An Overview of MetaMap: Historical Perspectives and recent advances. J Am Med Inform Assoc. 2010;17:229–36.
14. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34(5):301–10.
15. http://incubator.apache.org/opennlp/index.html.
16. http://www.ibm.com/watson.
17. http://uima.apache.org/documentation.html.
18. http://www.apelon.com/Products/DTS/tabid/97/Default.aspx.
19. Elkin PL, Tuttle M, Trusko B, Brown SH. Bioprospecting: Novel marker discover obtained from the bibleome. BMC Bioinformatics. 2009;10 Suppl 2:S9.
20. Elkin PL, Trusko BE, Koppel R, Speroff T, Mohrer D, Sakji S, Gurewitz I, Tuttle M, Brown SH. Secondary use of clinical data. Stud Health Technol Inform. 2010;155:14–29.
21. Elkin PL, Carter JS, Nabar M, Tuttle M, Lincoln M, Brown SH. Drug knowledge expressed as computable semantic triples. Stud Health Technol Inform. 2011;166:38–47.

Conclusion

10

Peter L. Elkin and Mark Samuel Tuttle

This textbook has been designed to teach students at all learning levels. This text has appropriate material for the student taking their first course in Biomedical Informatics to the advanced learner looking to take their career in healthcare terminologies to the next level.

We have learned the history of terminologies from ancient times to the present day successes and challenges, in health informatics terminological construction and application. The text addressed knowledge representation theory and its application in healthcare. We proceeded to expose the student to the theoretical foundations of terminology. We defined compositionality in healthcare and then exposed the student to the world of standards development in healthcare terminologies and encouraged their involvement. We discussed some of the major terminologies implemented for use in healthcare, and we defined some of the major terminological systems in use today as examples of the categories of terminological systems required in today's complex and ever evolving Health Informatics landscape.

The questions associated with each chapter are representative of the types of questions that one will need to know to function as a health terminologist and to develop a broad competency in Health Informatics. They are intended as a learning tool and provide a separate method for learning some of the content contained in the actual chapters.

In the introduction, we stated that this book was use case driven, and here we wish to make that connection clearly. Many use cases drive terminological development, dissemination, and implementation. However, a few of the use cases require all of the armamentarium that we have discussed in this textbook. Our use case involving a family involved in a car accident where the father's clinical records need to be transferred to his admitting hospital and the data and information in his record need to be used by the clinical decision support system at the admitting hospital, this use case drives most all of the requirements that we have for the development, distribution, and implementation of healthcare terminologies. When we couple this scenario with his son having a polymorphism in his CYP3A4 enzyme gene, we link in the new biology and personalized medicine which have also been a new driver of the need for accurate and consistent healthcare terminologies. Most other secondary uses of healthcare data require no greater depth of specification or level of accuracy than these clinical use cases.

What Have We Learned from Our Case?

That our patients deserve the highest quality and safest care that we can provide. This requires the use of all of the patient's relevant data codified by

P.L. Elkin, M.D., MACP, FACMI (✉)
Physician, Researcher and Author, 212 East 95th Street, Suite 3B, New York, NY 10128, USA
e-mail: ontolimatics@gmail.com

M.S. Tuttle, AB, BE, FACMI
Apelon, Ridgefield, CT, USA

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8_10,

high-quality healthcare terminologies meeting the highest level of standards and providing the greatest level of interoperability. We must utilize this codified data in the context of best practice of health and healthcare. Our hypothetical Mr. Kneivel and his son Michael are representatives of the people who have put their trust in us to provide for them the very best care. This requires systems engineering that can help us to integrate and analyze patient data in order to provide clinicians with just-in-time point-of-care best practice advice, in support of their medical practice.

Thank you for using our textbook entitled *Terminology and Terminological Systems* in the subspecialty of Health Terminologies within the field of Health Informatics. The learner that has mastered the material contained in this textbook will be well prepared for a career in Health Terminologies. Those students whose concentration is within other subspecialties of Health Informatics should know when to refer a case to a Health Terminology subspecialist. Writing this textbook has been fun and exciting, and it is my sincere hope that you have enjoyed this textbook as you have assimilated the information in this subspecialty of Health Informatics that has taken scientists and historians centuries to compile. By imparting a sense of history to the field in addition to the didactic information used by Health Terminology practitioners, I hope that learners will gain a perspective on the field and in doing so a sense of belonging to a tradition of excellence that has greatly contributed to our field of Health Informatics.

Index

P.L. Elkin (ed.), *Terminology and Terminological Systems*, Health Informatics,
DOI 10.1007/978-1-4471-2816-8, © Springer-Verlag London 2012