

Supporting a Multi-hierarchical Classification in the Object-Oriented Paradigm¹

Jemal H. Abawajy

Michael A. Shepherd

Department of Mathematics, Statistics & Computing Science

Dalhousie University

Halifax, Nova Scotia, Canada B3H 3J5

shepherd@cs.dal.ca

This research and the resulting prototype system show that the object-oriented paradigm is an appropriate mechanism for supporting a complex, multi-hierarchical controlled vocabulary and the resulting classification when applied to a data base. In addition to supporting such a classification, it allows the searching and browsing of both the data base items and the assigned vocabulary terms.

1. INTRODUCTION

In 1985, the National Library of Medicine initiated a Unified Medical Language System (UMLS) project to facilitate the retrieval and integration of biomedical information from several standard biomedical vocabulary sources [Humphreys, 1989]. These standard vocabularies have been developed to deal with different kinds of biomedical information and to satisfy different purposes. For example, the Medical Subject Headings (MeSH) [National Library of Medicine, 1991] system was designed for indexing bibliographic data bases, the International Classification of Diseases, 9th edition, Clinical Modification (ICD-9-CM) [Commission on Professional and Hospital Activities, 1978] was designed for coding hospital charts for disease, procedure, cause and vaccinations, and the Systematized Nomenclature of Medicine (SNOMED) [Cote, 1982] is a multi-axis classification scheme designed to encode clinical signs, symptoms, diagnosis, and procedures.

The UMLS consists of a Semantic Network, an Information Sources Map, and a Metathesaurus. The Metathesaurus [Tuttle, 1989, 1990] contains over 200,000 terms drawn from a number of different source vocabularies. Terms appearing in the Metathesaurus are categorized as main concept terms (preferred terms), synonyms of these main concept terms, related terms, or as lexical variants of either the main concept terms or of the synonyms. The Metathesaurus is hierarchical in nature so that core concepts are related to each other as broader, narrower, and other.

The integration of these vocabularies into a single metathesaurus brings with it many practical problems and inconsistencies. For instance, a simple hierarchy of concepts could be represented as a tree, i.e., a single root concept with each concept other than the root having at most one broader or superordinate concept and zero or more narrower or subordinate concepts. The Metathesaurus is, in fact, a directed graph as opposed to being a tree and has the following features:

1. This research was supported in part by the School of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada.

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

- **multiple hierarchies, each with its own root concept.** For example, “Morphology Axis” and “Functional Axis” in SNOMED, “Diseases and Injuries” in ICD, and “Diseases (MeSH Categories)” in MeSH are used to indicate the top most entry or root concepts in their respective hierarchies.
- **each concept, other than the root concepts, may have multiple immediate superordinate concepts (multiple parents).** This multi- hierarchical relationship is exemplified by examining the core concept “Sturge-Weber Syndrome” as it is presented in the Metathesaurus. This concept occurs within the MeSH hierarchy in four different places, as a hereditary neoplastic syndrome in two separate places, as an angiomatosis, and as a cause of mental retardation (see Figure 1). Note that all of these concepts are main or core concepts within the Metathesaurus.

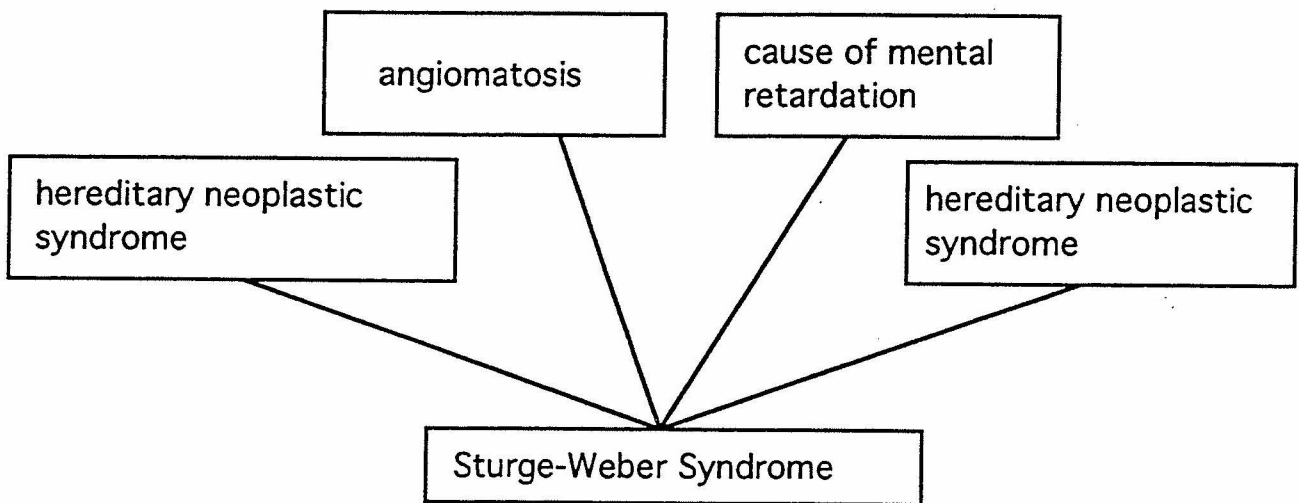


Figure 1. Multi-Hierarchical Relationship

- the same term (i.e., sequence of characters) may appear in completely different hierarchies.
- two concepts may be siblings in that they share a common superordinate concept while at the same time being in a superordinate-subordinate relationship to each other. For example, “Abdominal Wall” and “Respiratory Muscles” are both immediately subordinate to “Muscles”, yet “Abdominal Wall” and “Respiratory Muscles” are also in an immediate subordinate-superordinate relationship to each other (see Figure 2).

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

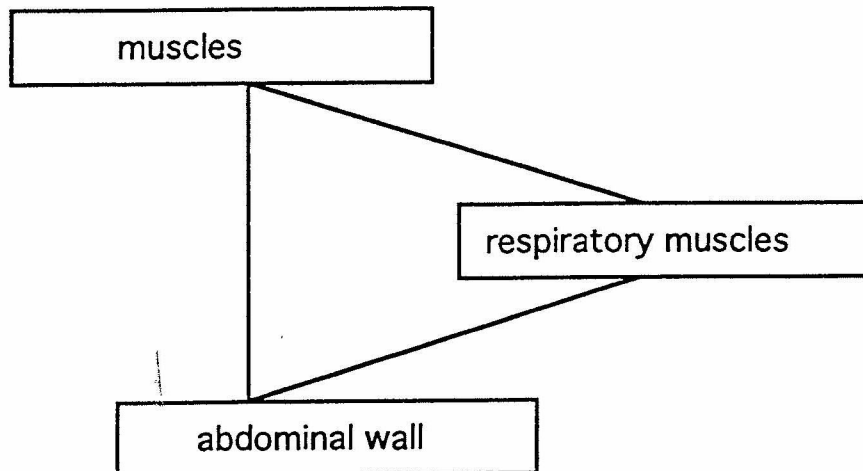


Figure 2. Simultaneous Parent-Child and Sibling Relationship

In the prototype system [Abawajy, 1992a, 1992b], the Metathesaurus was used to index automatically a medical curriculum data base of 1410 instructional units. Whenever a term from an instructional unit matched one or more main concept terms (or synonyms) in the Metathesaurus, the main concept term(s) and all of the terms in all of the hierarchies superordinate to those main concept terms were assigned to that instructional unit.

The prototype system is based on the object-oriented paradigm. Each assigned concept was defined as an object within this paradigm and the relationships between concepts were used to organize these objects into hierarchical classes of concept objects, resulting in a classification representing the medical curriculum.

Each instructional unit was also defined as an object. The classification of concepts was coupled with the data base of instructional units, thus producing hierarchical classes of instructional unit objects reflecting the concept classification. The instructional unit objects and the classification of concept objects were coupled but kept separate for access purposes. Thus, the classification of the information contained in the instructional units could be searched and browsed as a knowledge base, separately from the data base of instructional units itself.

Based on this research and the resulting prototype system, the object-oriented paradigm appears to be an appropriate mechanism for the support of complex relationships as exhibited in the UMLS Metathesaurus. The object-oriented paradigm supports single inheritance, multiple inheritance, and composite classes consisting of an aggregation or composition of nonhierarchically related

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

objects. This allows classifications with the complex features described above to be supported in a natural manner, consistent with the underlying software paradigm.

2. OVERVIEW OF THE METATHESAURUS

Version 1.1 of the Metathesaurus (Meta 1.1) contains over 200,000 terms drawn from MeSH, ICD-9-CM, SNOMED, Diagnostic and Statistical Manual of Mental Disorders (DSM), Current Procedural Terminology (CPT), LCSH, and from three different COSTAR sites. As the same concept may occur in more than one of the source vocabularies, one expression of the concept is taken to be the main concept or preferred term. The decision was made to avoid laborious selection among the alternative names for a given concept by establishing an order of precedence of source vocabularies. MESH has the highest precedence (assuming the concept is covered by MeSH) as it has both the broadest biomedical subject coverage and the most extensive contextual information for each of its terms.

There are over 43,000 reviewed main concepts together with a large number of lexical variations for these concepts. It includes information such as definitions, lexical category information, hierarchical contexts, and interrelationships among many of the concepts. It also includes other terms which have some relationships to the main concepts including synonyms, lexical variants, and related terms. In this paper, we are concerned only with the main concepts, synonyms, related terms, and the hierarchical contexts of the main concepts.

2.1. Core Concepts

The Metathesaurus is made up of multiple source vocabularies, each of which may consist of multiple hierarchies of concepts. For example, "Morphology Axis" and "Functional Axis" in SNOMED, "Diseases and Injuries" in ICD, and "Diseases (MeSH Categories)" in MeSH are used to indicate the top most entry in their respective hierarchies.

A main or core concept may associate hierarchically with more than one ancestor. This multi-hierarchical relationship is exemplified by examining the core concept "Sturge-Weber Syndrome" as it is presented in the Metathesaurus. This concept occurs within the MeSH hierarchy in four different places, as a hereditary neoplastic syndrome in two separate places, as an angiomatosis, and as a cause of mental retardation (see Figure 1). Note that all of these concepts are main or core concepts within the Metathesaurus.

In addition, two core concepts may be siblings in that they share a common superordinate concept while at the same time being in a superordinate-subordinate relationship to each other. For example, "Abdominal Wall" and "Respiratory Muscles" are both immediately subordinate to "Muscles", yet "Abdominal Wall" and "Respiratory Muscles" are also in an immediate subordinate-superordinate relationship to each other (see Figure 2).

2.2. Related Terms

The nomenclature, "Related Terms", is used to denote terms that have some relationship to the main concept name or one of its lexical variants or synonyms, but are not themselves main concept names. These terms are either indicated in one of the source vocabularies as being related to the main concept or its synonyms or are identified by lexic matching techniques and then determined by subject experts to be related terms. For example, there are 10 synonyms of the main concept term, "Abnormalities" (see Figure 7).

Such terms may be related to core concepts in broader, narrower, or "other" relationships. For example, "Cancer Care Units" is not a core or main concept, but is related to the core concept, "Cancer", in an "other" relationship and in a "narrower" relationship to the core concept, "Hospital Units".

2.3. Synonym Terms

Terms that have the same meaning as Core or Related terms, but are lexical variants of neither the Core terms nor the Related terms are called Synonyms. These terms are either explicitly linked to the Main concept term in one or more of the source vocabularies or identified by the lexical techniques used in developing the Metathesaurus.

The synonym relationship provides additional entry points into the core concept entries and provides a form of term that may be needed to retrieve information indexed by a particular vocabulary.

3. THE OBJECT-ORIENTED PARADIGM

In this section, we introduce some of the principal terms and concepts of the object-oriented paradigm that are pertinent to this paper, including classes, hierarchies, composite classes, and objects [Date, 1990; Ege, 1991].

3.1. Classes

In the object-oriented paradigm, the term "class" corresponds to the traditional programming notion of an abstract data type. The user can define their own classes which have attributes (called instance variables) and operators (called methods) which occur in every instance or object of that class. The only way to operate on an object is by means of the methods defined for that class.

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

3.2. Hierarchy

Classes are organized hierarchically in supertype/subtype fashion with subtype classes inheriting the instance variables and methods of their supertype classes. Therefore, a class may have instance variables and/or methods defined for that class, as well as instance variables and/or methods inherited from superordinate classes. This type of inter-class relationship is also called an is-a relationship. We say that A is-a B, read "A is a B", if class B is a generalization of class A.

3.3. Composite Class

Another type of inter-class relation is the aggregation or composition of objects to form a composite class. This is instantiated by an instance variable of an object of one class referring to another object of the same or another class. This type of relation is sometimes called a has-a relation and there is no inheritance involved. As shown in Figure 3, the Core class has-a parent, has-a synonym, has-a child, etc. Thus the Core class is a composite of these other five classes.

3.4. Objects

From a data modelling perspective, an object can be a thing, an organization, a person, or an event about which users wish to collect and store information. An object is an instance or member of a class, which defines its instance variables and methods. For instance, the term, "Abnormalities", is an object or instance of the Core class, while "Congenital Abnormalities" is an object or instance of the Synonym class.

Therefore, an object is an entity that has a state and behaviour. The state is represented by the values of the instance variables for that particular object. The behaviour of an object is specified by the methods of the class. Methods consist of code that modify or return the state of an object. Hence, methods are the only means of retrieving and updating the state of an object.

Upon instantiation, an object is assigned a unique identifier by the system. This identifier also identifies this object as a member of a specific class.

4. INDEXING THE CURRICULUM DATA BASE

Each instructional unit of the curriculum data base has a set of manually assigned descriptors. Each descriptor for an instructional unit was passed against the Metathesaurus. For each descriptor for which an entry in the Metathesaurus was found, the Metathesaurus term and all of its superordinate terms, synonym terms, and related terms were assigned to the instructional unit indexed by the original descriptor.

Classes and inter-class relationships were defined for both the terms assigned from the Metathesaurus and for the instructional units. Only the classes and relationships pertaining to

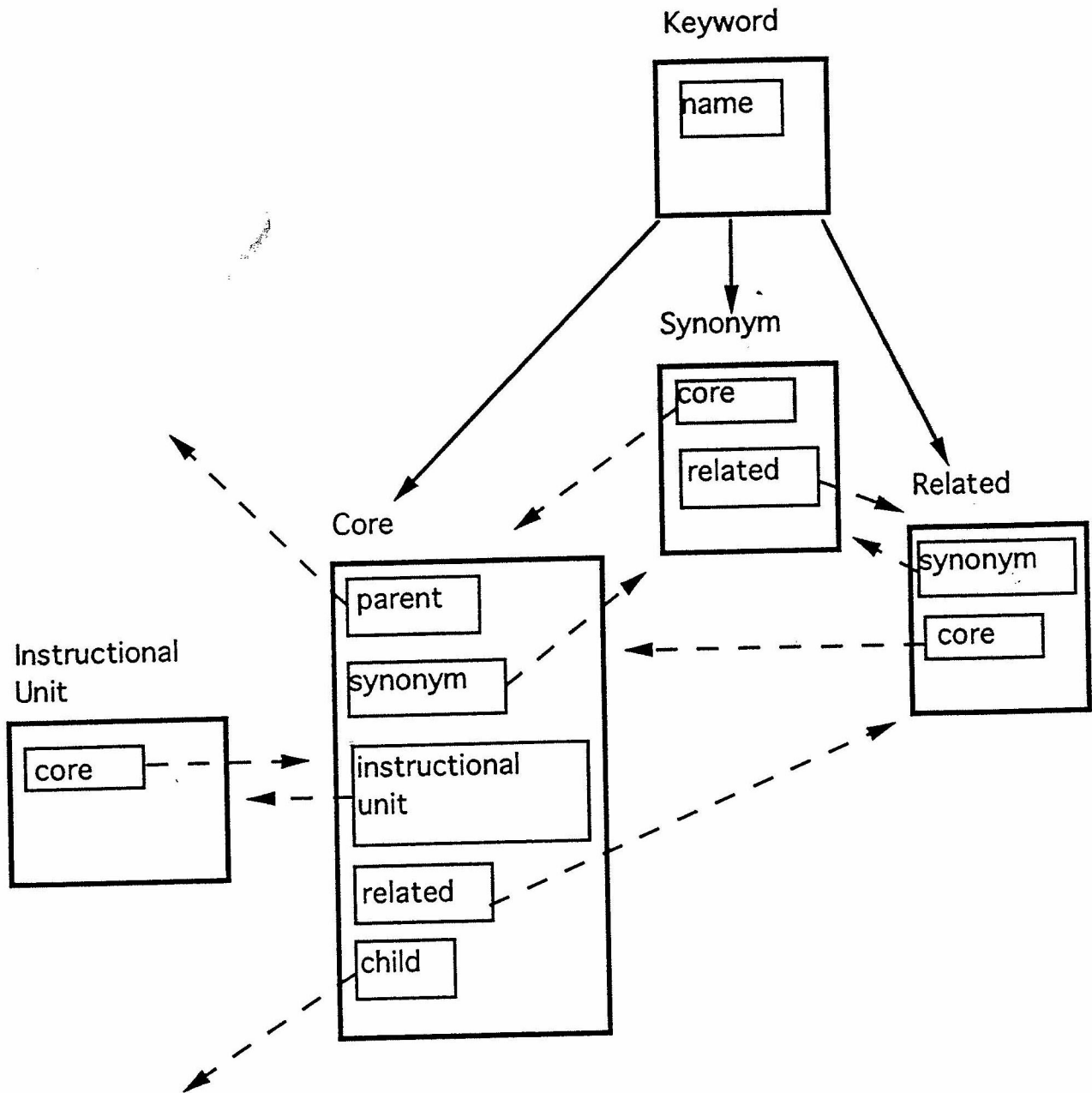


Figure 3. Class Definitions Assigned Metathesaurus Terms

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

the Metathesaurus are discussed in this paper. The classes pertaining to the instructional units making up the data base are not discussed.

The Keyword class describes the indexing language for the curriculum database. As shown in Figure 3, the Keyword class has an instance variable, Name, which is a character string used to identify each instance of the class by name. The Keyword class also has three subclasses which inherit the Name variable; Core, Synonym, and Related. The solid lines indicate the inheritance hierarchy, while the dashed lines indicate composite classes. These three subclasses are discussed below.

4.1. The Core Class

The Core class describes the main concept terms in the Metathesaurus. The Name variable is inherited from the Keyword class and is instantiated when a Core object is instantiated.

The Core class has Related, Synonym, Parent, and Child classes as composite classes, as indicated by the dashed lines of Figure 3. Many instances of the Core class may be related to many instances of the Related class. Therefore, the relationship between the Core class and the Related class are many-to-many.

The relationships between the Core class and the Synonym class, however, is one-to-many. This is because an instance of a Core class may have several instances of the Synonym class.

A core concept may have more than one parent as well as several children. These are defined as composite class relationships as indicated by the dashed lines of Figure 3. The relationship between the instances of the Core concept is recursive many-to-many. An instance of the Core class does not have to have parent concepts or children concepts to exist in the database. Whether or not instances of Synonym or Related classes are present in the database does not have any bearing on the Core class instance creation.

The Core class also has the Instructional Unit class as a composite class (Figure 3). When a descriptor from an instructional unit is matched in the Metathesaurus, a chain of main concept terms is extracted from the Metathesaurus and is instantiated as a chain of Core objects, linked through the parent and child instance variables. An instance of the Instructional Unit class is created and linked to the instance of the Core class that generated the chain of Core objects. The Instructional Unit Class is itself a complex object but is not discussed in this paper.

4.2. The Synonym Class

The Synonym class inherits the Name variable from the Keyword class and is instantiated when a Synonym object is instantiated.

The Synonym class has the Core and Related classes as composite classes (Figure 3).

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

Many instances of the Synonym class may be related to a core concept. Similarly, many instances of the Synonym class may be synonyms of a Related class instance. Therefore, the relationship between the Core class and the Synonym class as well as between the Related class and Synonym class is one-to-many. A synonym instance must have at least one instance of a related Core or Related class present in the database before it can be created.

4.3. The Related Class

The Related class inherits the Name variable from the Keyword class and is instantiated when a Related object is instantiated.

The Related class has the Core and Synonym classes as composite classes. Many instances of the Related class may relate to many instances of the Core class. Therefore, the relationship between the Related class and the Core class is many-to-many. The relationship between the Related class and the Synonym class, however, is one-to-many.

The association between the Related class and the Core class as well as between the Related class and the Synonym class is optional. Whether or not instances of Synonym and Core classes are present in the database does not have any bearing on the existence of instances of the Related class in the database.

5. THE PROTOTYPE SYSTEM

The system was written in C, an object-oriented programming language. Both the resulting concept classification (knowledge base) and the instructional unit data base were stored in ONTOS, an object-oriented data base management system. The user interface was developed using XView on a SUN workstation.

We have developed a browsing tool which enables the user to point and click to retrieve required information on the curriculum database or to display a biomedical concept, its definition, its synonyms, its related terms, semantic types, associated terms, or contexts as specified by the user.

The system permits the retrieval of an instructional unit in a number of different ways. In keeping with the scope of this paper, only retrieval through index terms is described and such features as retrieving faculty, methods of delivery, resource materials, etc., are not described.

5.1. Retrieving an Instructional Unit

In order to perform a search, the user selects a letter "A" through "Z". All core terms, synonym terms, and related terms beginning with the selected letter which were assigned to data base items will be displayed in the lower right-hand window. The user scrolls this window, browsing these terms, and clicks on an appropriate term. In Figure 4, the user has selected "Abdominal Wall" as

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

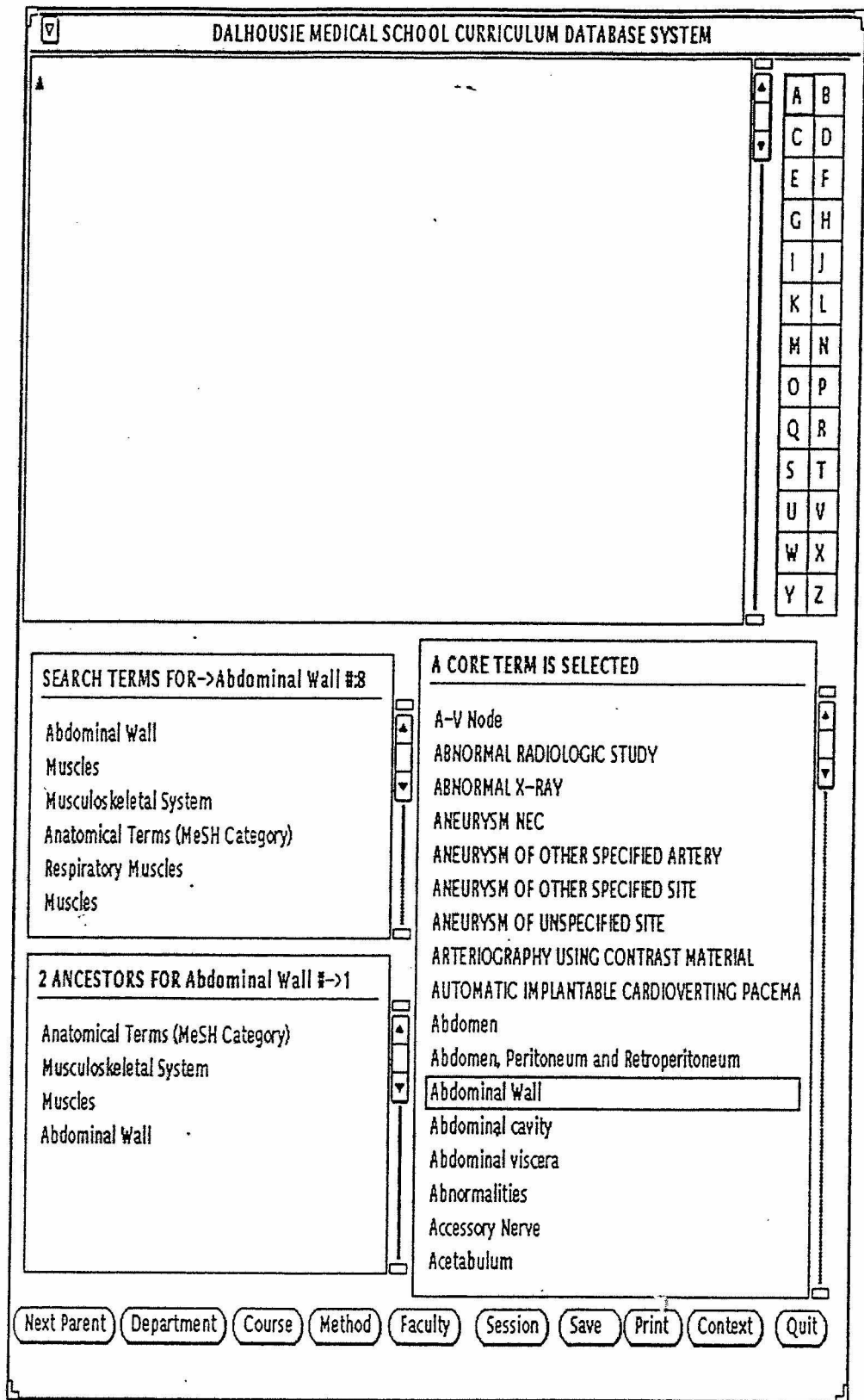


Figure 4. Selecting an Initial Search Term

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

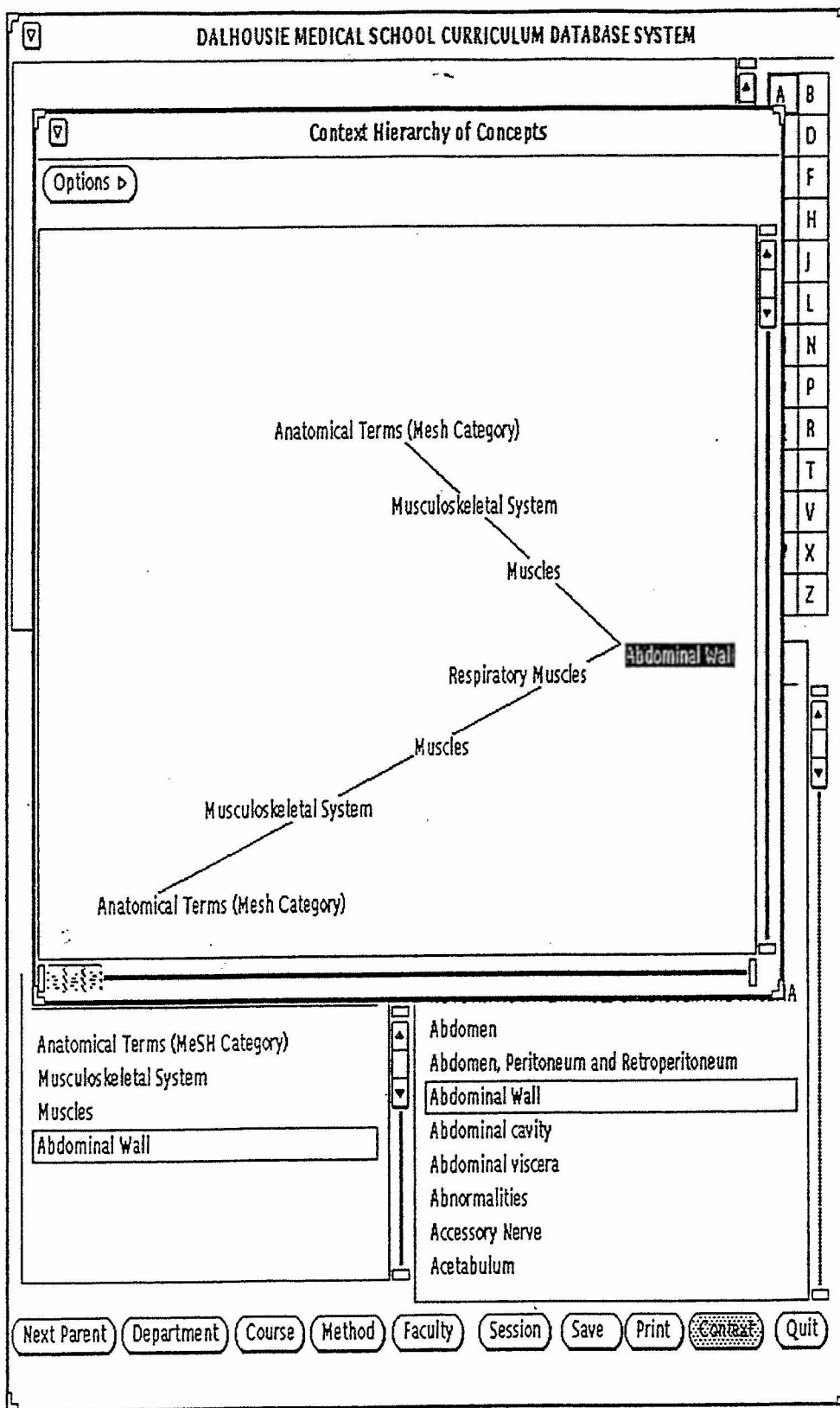


Figure 5. Graphical Representation of Ancestor Chains

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

DALHOUSIE MEDICAL SCHOOL CURRICULUM DATABASE SYSTEM

COURSE :Anatomy I
 TITLE :Gross - Anterior Abdominal Wall
 DATE :88/11/04
 METHOD :Lab
 DEPARTMENT :Anatomy department
 INSTRUCTOR :Dr. G. Sinha
 OBJECTIVE :To dissect and understand the external anatomy of the heart
 DESCRIPTION :Dissection of the anterior abdominal wall was completed. Dissection of the inguinal region was done to expose: superficial inguinal ring, inguinal ligament, spermatic cord - coverings and contents; ilioinguinal nerve; conjoint tendon; deep inguinal ring.
 REFERENCE :See appropriate sections of recommended texts and atlases

A	B
C	D
E	F
G	H
I	J
K	L
M	N
O	P
Q	R
S	T
U	V
W	X
Y	Z

FOUND 6 SESSION TITLES

- Gross - Anterior Abdominal Wall
- Gross - Inguinal Region
- Gross - Posterior Abdomen Wall, Kidney and Ur
- Gross Lab - Anterior Abdominal Wall
- Gross Lab: Inguinal Region
- Gross Lab: Posterior abdominal wall, kidney an

A CORE TERM IS SELECTED

- A-V Node
- ABNORMAL RADIOLOGIC STUDY
- ABNORMAL X-RAY
- ANEURYSM NEC
- ANEURYSM OF OTHER SPECIFIED ARTERY
- ANEURYSM OF OTHER SPECIFIED SITE
- ANEURYSM OF UNSPECIFIED SITE
- ARTERIOGRAPHY USING CONTRAST MATERIAL
- AUTOMATIC IMPLANTABLE CARDIOVERTING PACEMA
- Abdomen
- Abdomen, Peritoneum and Retroperitoneum
- Abdominal Wall
- Abdominal cavity
- Abdominal viscera
- Abnormalities
- Accessory Nerve
- Acetabulum

2 ANCESTORS FOR Abdominal Wall #->2

- Anatomical Terms (MeSH Category)
- Musculoskeletal System
- Muscles
- Respiratory Muscles
- Abdominal Wall

Next Parent
Department
Course
Method
Faculty
Session
Save
Print
Context
Quit

Figure 6. Retrieved Instructional Unit

the search term. The message in the title line of this window indicates that this is a core term (as opposed to a related term or synonym).

The selected term, "Abdominal Wall", and one of its core concept ancestor chains are displayed in the lower left window of Figure 4. This window indicates that there are two sets of core ancestor terms and that the first set is currently displayed. The list of possible search terms from which the user can make a selection has been expanded to include all core ancestor terms (i.e., from both ancestor chains) for a total of eight search terms.

Note that if the term selected originally had been a related term or a synonym term, an appropriate message would have been displayed in the title line of the lower right window, and the appropriate core term and the core terms ancestors would be displayed in the lower left window.

Figure 5 gives a graphical representation of the two chains of ancestors of "Abdominal Wall". Note that these two chains should come together at the common core concept, "Muscles", but do not (they will in the next version). By clicking on the various nodes, the user can move around the hierarchy and prune the list of search terms.

Figure 6 shows a completed data base search for all instructional units indexed by the core concept, "Abdominal Wall". Note that the six instructional units retrieved represent all units indexed directly by the term "Abdominal Wall" or indexed by a term having "Abdominal Wall" in that term's ancestor chain of core concepts. In this latter situation, "Abdominal Wall" is inferred to be an index term through the core concept hierarchy. One of the sessions has been retrieved and is displayed in the top window.

5.2. Browsing the Classification Scheme

As shown in Figure 7, the user may request to see the synonyms, related terms, siblings, and children, etc., of the selected core term. Figure 7 shows that there are ten synonym concepts for the core term, "Abnormalities".

Figure 8 shows that there are five terms related to the core term, "Abnormalities". Note that "Abnormalities" is in an "other" relationship with four of these related terms, while in a "broader" relationship with "Deformities".

Figure 9 shows that "Abdominal Wall" has six siblings. Note that the sibling relationship is not stored in the Metathesaurus, but is calculated dynamically, i.e., if two concepts have the same parent concept, then they are siblings. In this example, "Respiratory Muscles" is shown as a sibling in the Sibling Window, but as an ancestor in the Ancestor Window.

5.3. Changing the Scope of the Search

The scope of a search can be either broadened or narrowed. As one moves up the hierarchy, the search terms become more general and the scope of the resulting search is broadened. For instance,

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

Knowledge is represented by means of classes and relationships among classes, and each concept is an instantiation of a class. Applying object-oriented principles to the biomedical concepts with relationships as depicted in the Metathesaurus means to build up core, related, synonym, semantic, and lexical classes and to interconnect them via generalized/specialized class hierarchies and composite classes. Through this process the hierarchical nature of the biomedical concepts was defined succinctly.

The prototype system enables the user to select a phrase or a word from a unified biomedical controlled vocabulary and display its source vocabulary, definition, synonyms, related terms, semantic types, associated terms, or contexts as specified by the user. The users can query the system to find all concepts that have a particular characteristic, such as semantic type.

The number of biomedical terms is large and the semantic relationships among these terms are complex. As a result, the traditional data models can not easily capture the complex relationship between the biomedical concepts. The object-oriented model offers superior means of capturing

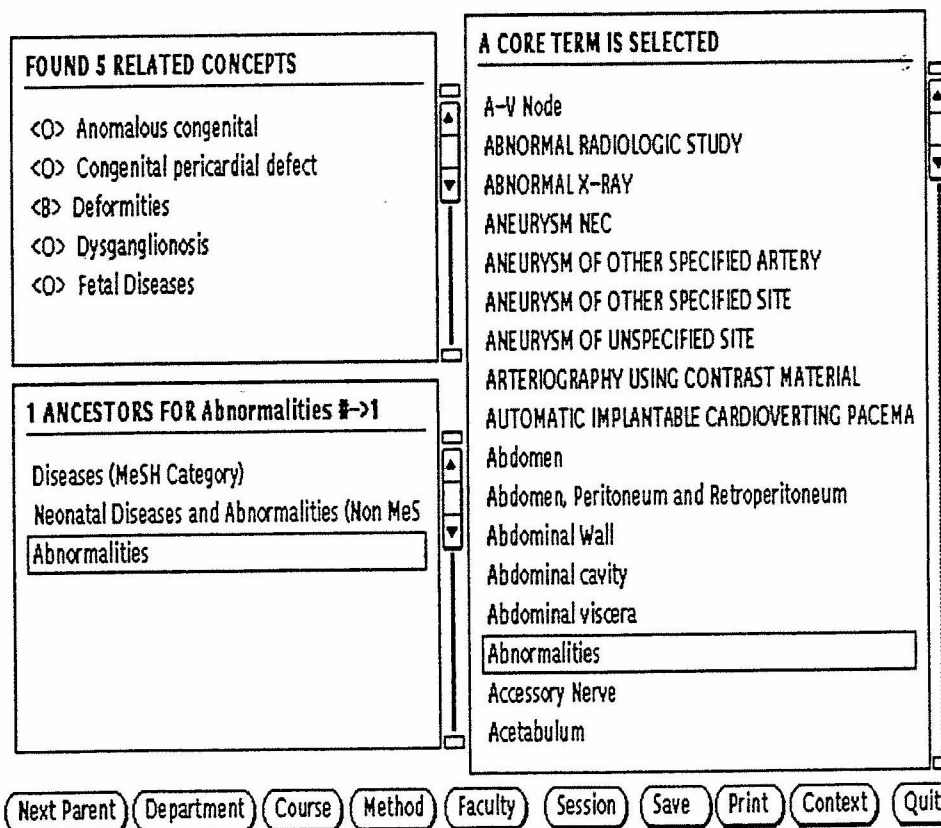


Figure 8. Display of Related Terms

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

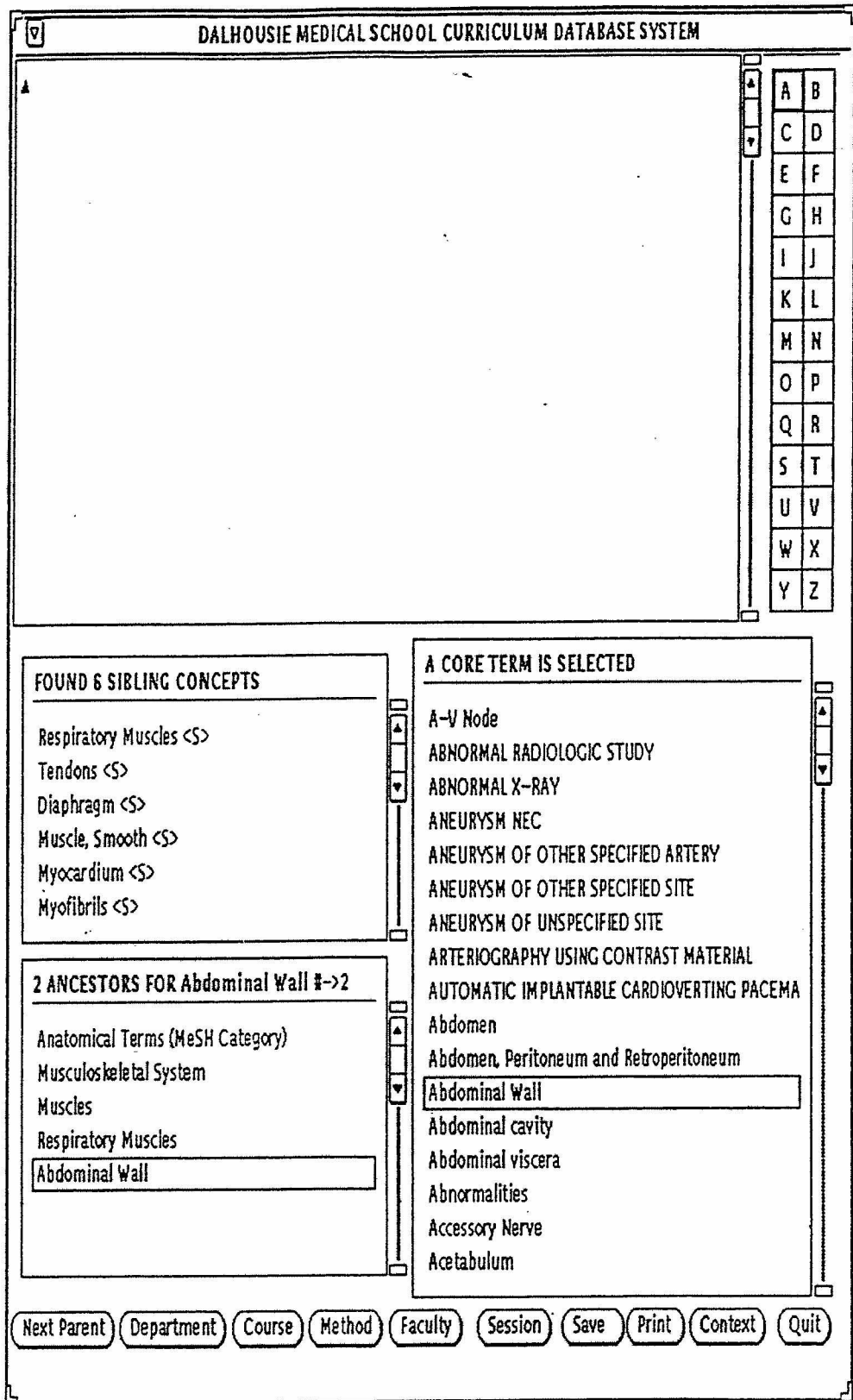


Figure 9. Display of Siblings

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

these complex relationships. A single entity is modelled as a single object, not as multiple tuples in multiple relations. Its fundamental components (such as classes which are used as templates to group objects of the same structure and behaviour) already have wide-spread applicability in modern information management systems.

7. REFERENCES

- Abawajy, Jemal H. 1992a. Knowledge Structuring With Unified Medical Language System. *Proceedings of the APICS Computer Science Conference*, Dalhousie University, Halifax, Nova Scotia, Canada. August 17.
- Abawajy, Jemal H. 1992b. *Design of an Integrated Knowledge/Data Base Medical Curriculum System: An Object-Oriented Framework and Experimental Prototype*. M.Sc. Thesis. Computing Science Division, Department of Mathematics, Statistics & Computing Science. Dalhousie University, Halifax, Nova Scotia, Canada.
- Commission on Professional and Hospital Activities. 1978. *International Classification of Diseases, 9th edition*, Clinical Modification. Commission on Professional and Hospital Activities. Ann Arbor, Michigan.
- Cote, R.A. 1982. *Systematic Nomenclature of Medicine*. 1982 College of American Pathologists. Skokie, Illinois.
- Date, C.J. 1990. *An Introduction to Database Systems, Volume 1, Fifth Edition*. Addison-Wesley: New York.
- Ege, Raimund K. 1991. *Programming in an Object-Oriented Environment*. Academic Press: New York.
- Humphreys, Betsy L. et al. 1989. Building the Unified Medical Language System. *SCAMC 1989 Proceedings*, Nov. 5-8, Washington, D.C., pp. 475-480.
- National Library of Medicine. 1991. *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland.
- Tuttle, M.S. et al. 1990. Using Meta-1 — the 1st Version of the UMLS Metathesaurus. *SCAMC 1990 Proceedings*, Nov. 4-7, Washington, D.C., pp. 131-135.
- Tuttle, M.S. et al. 1989. Implementing Meta-1 — the 1st Version of the UMLS. *SCAMC 1989 Proceedings*, Nov. 5-8, Washington, D.C., pp. 483-487.

PROCEEDINGS OF THE 4th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP