The Indexing Initiative

A Report to the Board of Scientific Counselors
of the Lister Hill National Center
for Biomedical Communications

October 14, 1999

Alan R. Aronson (LHNCBC), Olivier Bodenreider (LHNCBC),
H. Florence Chang (LHNCBC), Susanne M. Humphrey (LHNCBC),
James G. Mork (LHNCBC), Stuart J. Nelson (LO),
Thomas C. Rindflesch (LHNCBC), W. John Wilbur (NCBI)

Table of Contents

1. Background 1
2. Project Objectives 1
3. Project Significance 1
4. Methods and Procedures 2
   - 4.1 The IND Prototype 2
     - 4.1.1 MetaMap Indexing 3
     - 4.1.2 Barrier Words with Approximate Matching 6
     - 4.1.3 Trigrams 7
     - 4.1.4 Restrict to MeSH 8
     - 4.1.5 PubMed Related Citations 12
     - 4.1.6 The INQUERY Method 13
     - 4.1.7 Clustering 14
   - 4.2 Experiments 15
   - 4.3 An Example 17
   - 4.4 Journal Descriptor (JD) Indexing 19
5. Evaluation 23
   - 5.1 Background 23
   - 5.2 Index-based Evaluation 24
   - 5.3 Retrieval-based Evaluation 25
   - 5.4 User-centered Evaluation 26
6. Project Plan 26
   - 6.1 Technical Development of the Prototype 26
   - 6.2 Enhancements to the Prototype Based on Results of Evaluation 27
     - 6.2.1 Word Sense Disambiguation 27
     - 6.2.2 Full Text Processing 27
     - 6.2.3 Semantic Proximity 27
   - 6.3 Enhancements Based on Journal Descriptor Indexing 28
   - 6.4 Enhancements Based on Machine Learning 28
7. Summary 28
8. References 29
9. Appendix: Indexing Initiative Team Members 32

1. Background

For more than 150 years, the National Library of Medicine has provided access to the biomedical journal literature through the analytical efforts of human indexers. Since 1966, access has been provided in the form of electronically searchable document surrogates consisting of bibliographic citations, descriptors assigned by indexers from the MeSH® controlled vocabulary (MeSH, 1998) and, since 1974, author abstracts of many, but not all, items.

In the late 1990s, as medical journals migrate from print to electronic form, the need for human intervention to link users with relevant documents may be minimized, if not eliminated altogether. In addition, the cost of human indexing of the biomedical literature is high. As budgets are reduced and costs continue to climb, it seems reasonable to investigate alternative methods for indexing bibliographic and other data.

The MEDLINE® database contains about 11 million records, all of which have been produced by human indexing. The file presently grows at the rate of about 400,000 indexed citations per year, covering about 4,300 international biomedical journals. Human indexing consists of reviewing the complete text of each article, rather than an abstract or summary of it, and assigning descriptors that represent the central concepts as well as every other topic that is discussed to a significant extent. Indexers assign descriptors from the MeSH vocabulary of more than 19,000 main headings. Main heading descriptors may be further qualified by selections from a collection of 88 topical subheadings.

Since 1990, there has been a steady and sizeable increase in the number of articles received, owing both to an increase in the number of indexed journals and, to a lesser extent, to an increase in the number of articles in journals that are already being indexed.

In the face of a growing workload and dwindling resources, we have undertaken the Indexing Initiative to re-examine both the way that MEDLINE is currently produced and also the ways in which NLM might accomplish its mission of providing access to biomedical literature other than by manual subject indexing.

2. Project Objectives

The objective of NLM's Indexing Initiative (IND) is to investigate methods whereby automated indexing methods partially or completely substitute for current indexing practices. The project will be considered a success if methods can be designed and implemented that result in retrieval performance that is equal to or better than the retrieval performance of systems based principally on humanly assigned index terms.

3. Project Significance

Human indexing is an expensive and labor-intensive activity. The total costs of indexing at NLM include data entry, NLM staff indexing and revising, contract indexing, equipment, and telecommunications costs. Indexers are highly trained individuals, not only in MEDLINE indexing practice, but also in one or several of the subject domains covered by the MEDLINE database. It is becoming increasingly difficult to hire indexers with the level of expertise that is necessary for indexing the scientific literature in MEDLINE.

All of these considerations indicate that if automated methods can be successfully developed and implemented, the project will have an important impact on NLM's ability to continue to provide high-quality services to its constituents. Secondarily, but also importantly, the project should contribute significantly to information science research. As more and more documents become available in electronic form, and as more and more organizations develop "digital libraries" for their collections, automated techniques for accessing the information are required. It is not possible to index each document by hand, and new methods must be developed.

4. Methods and Procedures

The project will assume the continued existence and growth of NLM's MeSH vocabulary and of the UMLS® Knowledge Sources (Lindberg, Humphreys, and McCray, 1993a; UMLS, 1998). It will assume the availability of free text in the form of titles and abstracts but will also consider the increasing availability of the full text of journal articles in electronic form. The project will investigate concept-based indexing methods that go well beyond automatic word-based indexing (such as the inverted word index already part of MEDLINE). The technical plan for the Indexing Initiative involves the design of experiments to test hypotheses such as the following.

1) Algorithms that exploit the UMLS Knowledge Sources will result in a set of usable MeSH terms and relationships for characterizing MEDLINE titles and abstracts.

2) Automated techniques for indexing MEDLINE titles and abstracts will result in adequate retrieval performance when compared with retrieval from humanly indexed MEDLINE citations based on full text.

3) Algorithms that exploit clustering techniques will result in adequate retrieval performance when compared with retrieval from humanly indexed MEDLINE citations.

4) Automated indexing of full text results in superior retrieval performance when compared with indexing of titles and abstracts only.

5) Automated indexing techniques can be used to identify new terminology for inclusion in the MeSH thesaurus.

Recent IND efforts have focused on the creation of a system for exploring different ways of producing recommended indexing terms. Experiments using the system showed which combination of methods produced the best results. The best combination of methods has been incorporated into a version of the system called the IND Prototype.

4.1 The IND Prototype

The IND Prototype system consists of software for applying alternative methods of discovering MeSH headings for citation titles and abstracts and then combining them into an ordered list of recommended indexing terms as shown in Figure 1. The top portion of the diagram consists of five paths, or methods, for creating a list of recommended indexing terms: the INQUERY method, MetaMap Indexing, Barrier Words with Approximate Matching, Trigrams, and PubMed Related Citations. The middle three paths actually compute UMLS Metathesaurus® concepts which are passed to the Restrict to MeSH method, and the outer two paths compute MeSH headings directly. The results from each path are weighted and combined using the Clustering method. The system

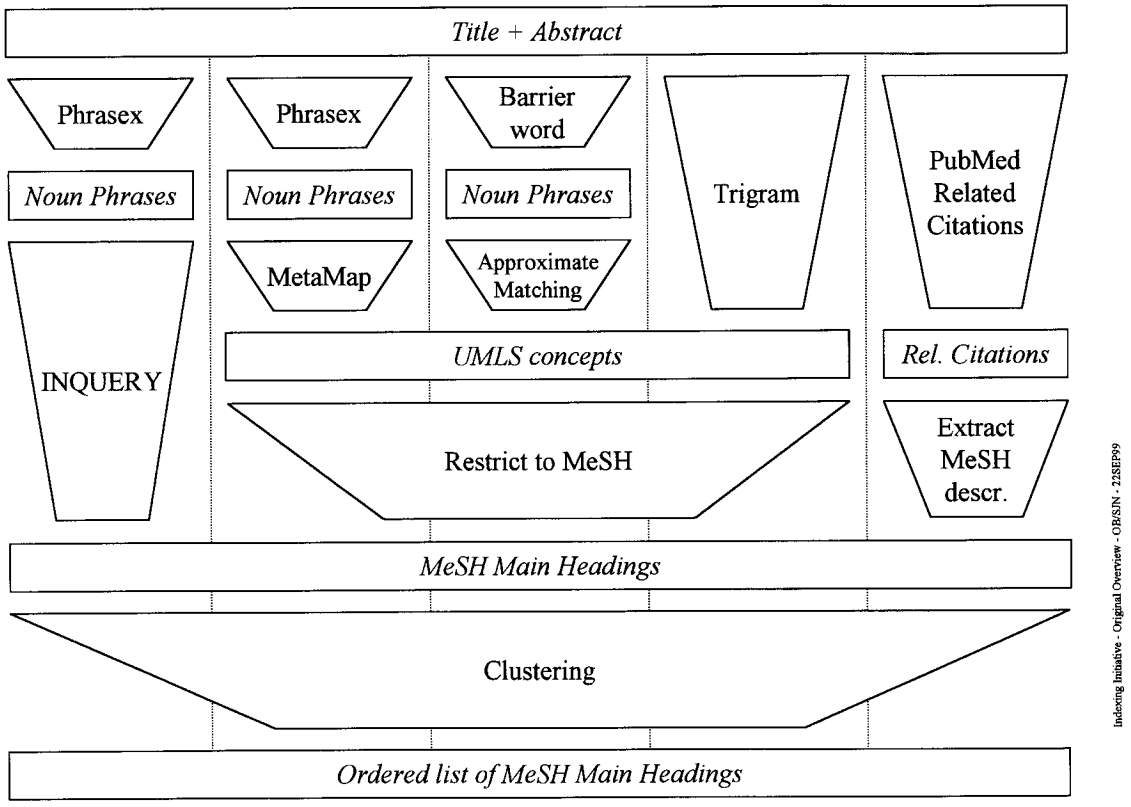

Figure 1. The Indexing Initiative System

is highly parameterized not only by path weights but also by several parameters specific to the Restrict to MeSH and Clustering methods. The remainder of this section contains descriptions of all of the methods and a discussion of some experiments that determined the combination of paths producing the best results. The resulting optimized system is referred to as the Indexing Initiative Prototype, or simply the IND Prototype.

The order in which the methods will be described is as follows: the three methods which compute Metathesaurus concepts (MetaMap Indexing, Barrier Words with Approximate Matching, and Trigrams), the Restrict to MeSH method, the two methods which compute MeSH headings directly (PubMed Related Citations and the INQUERY method), and finally the Clustering method.

4.1.1 MetaMap Indexing

The MetaMap Indexing (MMI) method of discovering UMLS concepts consists of applying MetaMap (Aronson et al., 1994; Aronson and Rindflesch, 1997) to a body of text and then ordering the resulting concepts using a ranking function. Each process is described in turn.

MetaMap Processing

MetaMap performs the task of mapping from biomedical text to concepts in the UMLS Metathesaurus in the following five steps:

1. Parsing
   Arbitrary text is parsed into simple noun phrases; this limits the scope of further processing and thereby makes the mapping effort more tractable. Parsing is accomplished using the SPECIALIST™ minimal commitment parser (McCray et al., 1993) which produces a shallow syntactic analysis rather than a deep syntactic analysis. The parser normally uses the Xerox Part-of-speech tagger (Cutting et al., 1992) which assigns syntactic labels (e.g., noun, verb) to all textual items. The parser is good at determining the simple noun phrases in text, and the tagger improves accuracy even more.

   Consider the citation title *Bupivacaine inhibition of L-type calcium current in ventricular cardiomyocytes of hamster.* The parser detects four noun phrases: *Bupivacaine inhibition*, *L-type calcium current*, *ventricular cardiomyocytes*, and *hamster.* A simplified syntactic analysis for *Bupivacaine inhibition* is `[mod(bupivacaine), head(inhibition)]`. Note that the parser indicates that *inhibition* is the most central part, the *head*, of the phrase.

2. Variant Generation
   For each phrase, variants are generated where a variant consists of one or more consecutive phrase words (called a *generator*) together with all its acronyms, abbreviations, synonyms, derivational variants and meaningful combinations of these (Aronson, 1996). This is shown pictorially in Figure 2. The variants of the generator *inhibition* are shown in Figure 3. The variants

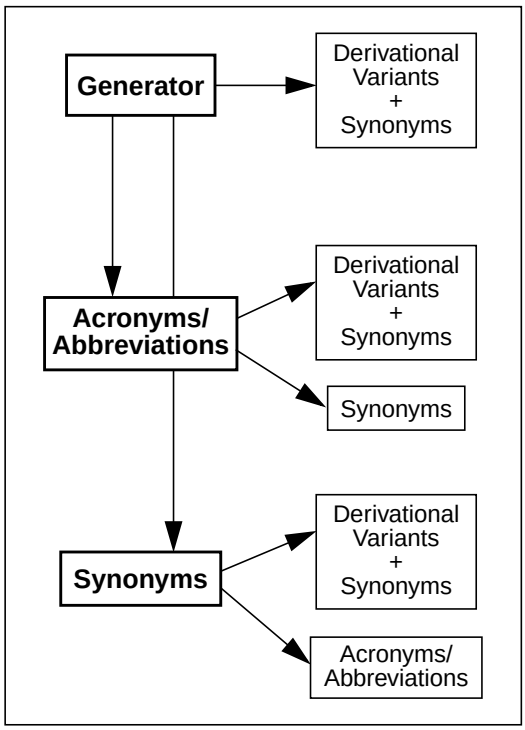

Figure 2. MetaMap variant generation

   are arranged hierarchically according to their derivation history. Each variant is followed by its part of speech, its distance score from its generator and its history. For example, *inhibition* (a noun) has distance score 0 and empty history because it is a generator. Similarly, the noun

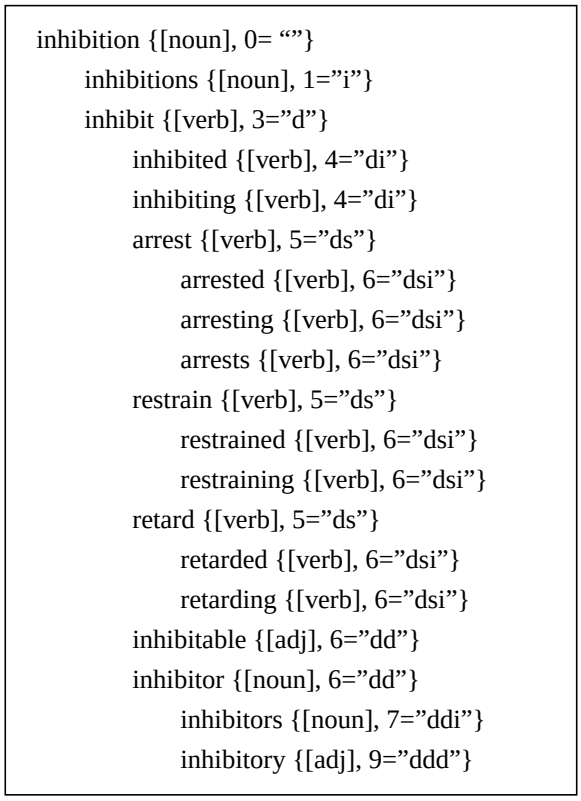

Figure 3. The variants of *inhibition*

inhibitors has distance score 7 and history “ddi” meaning that it is an inflection of a derivational variant (*inhibitor*) of a derivational variant (*inhibit*) of *inhibition*.

3. Candidate Retrieval
   The *candidate set* of all Metathesaurus strings containing at least one of the variants is retrieved.

4. Candidate Evaluation
   Each Metathesaurus candidate is evaluated against the input text by first computing a mapping from the phrase words to the candidate’s words and then calculating the strength of the mapping using a linguistically principled evaluation function consisting of a weighted average of four metrics: centrality (involvement of the head), variation, coverage and cohesiveness. The candidates are ordered according to mapping strength.

   The eight candidates for the phrase *Bupivacaine inhibition* are shown in Figure 4 ordered by mapping strength which has been normalized to a score between 0 and 1,000. If the candidate is not the preferred name of the concept it represents, the preferred name is displayed in parentheses. Note that all of the candidates corresponding to the text *inhibition* score better than “Bupivacaine” because they involve the head of the phrase.

5. Mapping Construction
   Complete mappings are constructed by combining candidates involved in disjoint parts of the phrase, and the strength of the complete mappings is computed just as for candidate mappings. The highest-scoring complete mappings represent MetaMap’s best interpretation of the original phrase.
   The highest ranked mappings for the phrase *Bupivacaine inhibition* consist of the Meta-

861 Inhibition <1> (Psychological inhibition) [Mental Process]

861 Inhibition, NOS (inhibition, physical) [Molecular Function]

777 Inhibited [Qualitative Concept]

768 Arrest (Arrested progression) [Temporal Concept]

768 arrest <1> (Law enforcement arrest) [Governmental or Regulatory Activity]

761 Retarded [Qualitative Concept]

755 inhibitors [Chemical Viewed Functionally]

694 Bupivacaine [Organic Chemical, Pharmacologic Substance]

Figure 4. Metathesaurus candidates for *Bupivacaine inhibition*

thesaurus concept “Bupivacaine” and either the concept “Psychological inhibition” or the concept “inhibition, physical”.

The MMI Ranking Function

The ranking function for MetaMap indexing is the product of a *frequency* factor and a *relevance* factor. The relevance factor is, in turn, a weighted average of four components (listed in order of importance): a MeSH tree depth factor, a word length factor, a character count factor, and a MetaMap score factor. For concepts found in the title of a MEDLINE citation, there is a simplified form of the function. Denoting MeSH tree depth, word count, character count, and MetaMap score, by *m, w, c* and *mm*, respectively, the ranking function can be written as:

$$\nu_5(\boldsymbol{f}) \cdot \frac{12 \cdot \nu_0(\boldsymbol{m}) + 2 \cdot \nu_{-10}(\boldsymbol{w}) + 2 \cdot \nu_{-10}(\boldsymbol{c}) + 1 \cdot \nu_{-10}(\boldsymbol{mm})}{12 + 2 + 2 + 1}$$

where the functions ν_n are normalizing sigmoidal functions. When the concept occurs in the title of the citation, the frequency factor $\nu_5(f)$ is replaced by 1. This has the effect of giving title concepts overwhelmingly good rankings. See Section 4.3 for a ranking example.

4.1.2 Barrier Words with Approximate Matching

The barrier word method is a fast way of identifying short noun phrases in free text. The text is parsed into sentences, where a sentence is computed as a set of words beginning with a capital letter and delimited by terminating punctuation. A potential nominal phrase is computed as a sequence of words occurring between *barrier words*, which are derived from a set of stopwords including articles, prepositions, and verbs. For example, consider the text: *The local anesthetic bupivacaine is cardiotoxic when accidentally injected into the circulation.* The set of barrier words might be used to identify *local anesthetic bupivacaine*, *cardiotoxic*, and *circulation* as nominal phrases. While this method has been used for some time, the use of a very long list of barrier words (approximately 24,000) was found to be much more effective in identifying nominal phrases in text than the traditional shorter lists (see Nelson et al., 1995).

Recent refinements and additions to the method include using a two-step parsing procedure as well as allowing certain words, which would otherwise be stopwords, to be included in the middle of the phrase. The two-step procedure first screens for chemical names (which otherwise might have punctuation marks or other linguistic infelicities in them). Chemical names occurring in the text are identified by a "sliding window technique" which requires the use of a long list of known chemicals. Such a list was obtained from the UMLS. Character by character, a string of characters is matched against this known list. As long as the string continues to match a chemical on the list, the process continues. It stops only when either the character by character match can no longer continue, or until a complete match with a chemical on the list occurs.

The second refinement of the method removes articles and prepositions, which might occur in the midst of a nominal phrase, from the barrier word list if the word preceding them is not in the barrier word list. Thus *carcinoma of the pancreas* would be identified as a nominal phrase despite the presence of the words *of* and *the*, which would otherwise be barrier words. The phrase *consists of the pancreas* would correctly not be identified as a nominal phrase because *consists*, a verb, is also a barrier word.

The noun phrases identified using the barrier word method are then processed by the Approximate Matching algorithm, a version of MetaMap's browse mode created for use in NLM's Large Scale Vocabulary Test. Approximate matching is much less strict than normal (semantic) MetaMap processing. In effect, it casts a wider net to locate Metathesaurus concepts that would be missed by MetaMap's semantic mode. This produces more concepts at the risk of including ones not closely related to the input noun phrase.

4.1.3 Trigrams

Trigram Phrase Matching is a method of identifying phrases that have a high probability of being synonyms. It is based on representing each phrase by a set of character trigrams that are extracted from that phrase. The character trigrams are used as key terms in a representation of the phrase much as words are used as key terms to represent a document. The similarity of phrases is then computed using the vector cosine similarity measure.

Trigram production rules

Let *STR* be the string of characters that represent a phrase. *STR* is processed as follows.

1) *STR* is lower cased.

2) *STR* is broken into terms at spaces and these individual terms are used to produce trigrams. Strings of length k+3 produce k+1 overlapping trigrams, while any string of length 3 or shorter is taken as the only trigram produced (for simplicity we shall refer to it as a trigram even if it has only one or two letters). All such trigrams are attributes of *STR*.

3) The first trigram produced from each term derived from *STR* is marked at the right end by the addition of the symbol '!' and the result is included as an attribute with a local count of 2. Also the first letter of the term is marked by adding the character '#' to the right and included as an attribute. Finally between any two adjacent terms in the phrase the trigram which consists of the first letters separated by a space is added as an attribute.

As an example consider the phrase “DNA sequence selectivity”. This is lower cased to “dna sequence selectivity” and step 2) gives rise to the trigrams dna, seq, equ, que, uen, enc, nce, sel, ele, lec, ect, cti, tiv, ivi, vit, ity. Step 3) then adds the attributes dna!, d#, seq!, s#, sel!, s#, d s, s s.

Attribute Weighting

All attributes are given global weights of the form $\sqrt{\log(N/n_t)}$. These are relatively standard inverse document frequency weights. Here N represents the collection size, which in this case is the number of phrases in the set we are studying. The value n_t is the frequency of occurrence throughout the collection of the attribute being weighted. Each attribute is also given a local weight which is $\log(1+f_t)$ where f_t is the number of times the attribute is seen in the particular phrase where the local weight is to be applied.

When all attributes have been weighted then each phrase is represented by the vector of local times global weights for all attributes as computed for that phrase. As usual only those attributes that actually occur in a phrase have nonzero coordinates in this vector representation.

The Similarity Score

The similarity between two phrases is then computed as the cosine of the angle between them. This is always a number between 0 and 1. We have found that when the score is roughly 0.7 or greater the probability becomes high that the two phrases are synonymous in meaning.

Application to Indexing

For purposes of indexing we process according to the following algorithm:

1) Break the title and abstract of a document up into all possible phrases consisting of one to six contiguous words without punctuation occurring within.

2) For each phrase produced in 1) compute the score against all phrases in UMLS and record the phrase that obtains the highest score.

3) For each word in the title and abstract, record that phrase of which that word is a member and which receives the highest overall score against the UMLS and record also the UMLS phrase that produced that highest score.

4) For each phrase pair obtained in 3) where one element is a phrase in the document and the other is a phrase in UMLS, count how many times the pair appears in different places in the document and return the pair, their score, and the count.

Like MetaMap Indexing and Barrier Words with Approximate Matching, this algorithm produces UMLS concepts which are subsequently restricted to MeSH headings as described in the next section.

4.1.4 Restrict to MeSH

Background

The representation of meaning in the UMLS is organized according to the principle of semantic locality (Nelson et al., 1991; McCray and Nelson, 1995), in which several means of representing relationships between concepts conspire to produce a cluster of semantically-related terms.

Dimensions of semantic locality include term information (synonymy, hypernymy, hyponymy), contextual information in a particular source vocabulary, co-occurrence of terms in the medical literature, and the categorization of concepts in the Semantic Network. In the Indexing Initiative, three of these phenomena are used to find the MeSH terms most closely related to any given UMLS concept: synonyms, interconcept relationships, and categorization (Bodenreider et al., 1998; also see Cimino et al., 1993).

In the UMLS, terms which are equivalent in meaning (synonyms) are clustered into a unique concept and are categorized with a semantic type. Furthermore, interconcept relationships, either inherited from the source vocabularies or specifically generated, provide additional semantic structure. For example, the SNOMED term “Gene-directed cell death” and the synonymous MeSH term “Apoptosis” are included in the same concept in the Metathesaurus, and this concept is given the semantic type ‘Cell Function.’ Within this same semantic locality, interconcept relations (both hierarchical and non-hierarchical) are seen in “Cell Death” as the parent of “Apoptosis,” as well as in “cell killing” and “Caspases” as other related concepts.

A further resource available in the Metathesaurus is also exploited by the algorithm for restricting concepts to MeSH. Although the UMLS is mostly a collection of precoordinated terms, the associated expressions created by indexers provide a translation of some complex concepts to expressions in other vocabularies. This is done by using elementary concepts combined with both logical operators and possibly, in mappings to MeSH, main heading and subheading combinations. For example the non-MeSH term “Mumps pancreatitis” has the associated expression “Mumps/complications AND Pancreatitis/etiology,” which in addition to the two MeSH main headings, includes the subheadings “complications” and “etiology.”

Methods

Based on the aspects of semantic locality noted, the overall strategy for restricting a given UMLS term to the semantically closest MeSH term involves the following four steps:

1. Choose a MeSH term as a synonym of the source concept.
2. Choose an associated expression which is a translation of the source concept.
3. Select MeSH terms from concepts hierarchically related to the source concept.
4. Base the selection on the non-hierarchically related concepts of the source concept.

The algorithm stops at any step that succeeds. Thus, if the source concept is “Gene-directed cell death,” step 1 of the algorithm chooses “Apoptosis” as the best MeSH equivalent. In the case of “Mumps pancreatitis,” which does not have a MeSH synonym, step 2 selects the associated expression for this term given above. Step 3, in which MeSH terms are chosen from the set of terms hierarchically related to the source concept (ancestors), is the most complex part of the algorithm and is described below.

Finally, if no MeSH term can be found from the ancestors, the non-hierarchically related concepts are explored in step 4. As noted above, these concepts are related to the source concept, but the exact nature of this relationship is not given in the Metathesaurus. For example, “Choroidal detachment, NOS” is related in this way to the MeSH term “Retinal Detachment.” In step 4, steps 1 to 3 are applied to each concept related non-hierarchically to the source concept.

Determining hierarchically-related concepts (Step 3)

In cases where a source concept does not map to a MeSH concept either as a synonym or as an associated expression, a set of MeSH terms is selected from the ancestors of the source concept as the best MeSH representative of the meaning of that term. For example, for the source concept "Vein of neck, NOS," step 3 in the algorithm chooses the two terms "Neck" and "Veins" as the closest MeSH equivalents. The algorithm to determine these hierarchically-related concepts builds a graph of ancestors using the source concept as the starting point, or seed. Then from this graph the MeSH terms closest to the starting point are selected.

Siblings and children as well as narrower concepts of the source concept can be used as the seed of the graph when no MeSH terms can be found in the graph started by the source concept. When concepts other than the source concept itself are used as the seed for the graph, the concepts chosen as the seed must be compatible in semantic type assignment.

Two concepts are defined as being compatible when they have identical semantic types or when at least one of the semantic types of the first concept is in an 'inverse_isa' relationship in the Semantic Network to at least one of the semantic types of the second concept. For example, the semantic type 'Anatomical Structure' is in an inverse_isa relationship with the semantic type 'Body Part, Organ, or Organ Component.'

Step 3 of the algorithm to choose the semantically best MeSH term for any given UMLS concept proceeds in two phases:

Phase 1: Build a graph of the ancestors of the source concept. The ancestors of a given concept can be represented as a directed graph, ideally acyclic. Starting from the seed (or source concept), its parents and broader concepts are added to the graph. Then, recursively, parents and broader concepts of all newly added concepts are added, until no new concept can be found.

The graph of the ancestors of the seed concept "Vein of Neck, NOS" is given in Figure 5, where the immediate ancestors of this concept are "Neck" and "Vein of head and neck, NOS." Ancestors of these terms include "Head" as a parent of "Neck" and both "Veins" and "Head" as broader concepts of "Vein of head and neck, NOS." Note that both MeSH ("Neck," "Head," and "Veins") and non-MeSH ("Vein of head and neck, NOS") ancestors are added to the graph. This insures that the MeSH concept "Veins" is included in the graph by virtue of being an immediate ancestor of non-MeSH "Vein of head and neck, NOS."

To prevent non relevant concepts from being added to the graph, the semantic types of any concept added to the graph must be compatible with those of its direct descendant in the graph. Once all ancestors, both MeSH and non-MeSH, have been added to the graph, phase 2 restricts the ancestors to those MeSH terms which best represent the meaning of the original source concept.

Phase 2: Select MeSH terms from the ancestors. All terms other than MeSH terms are first eliminated from the graph of ancestors. In Figure 5, this has the effect of removing all but five terms: "Neck," "Head," "Veins," "Body Regions," and "Blood Vessels."

Finally, MeSH candidates that are ancestors of other MeSH terms are removed. "Blood Vessels" is erased from the graph, since it is an ancestor of "Veins"; "Head" and "Body Regions" are also removed, being ancestors of "Neck." This procedure has the effect of insuring that the MeSH terms chosen as the final semantic representation of the source concept are the most specific terms

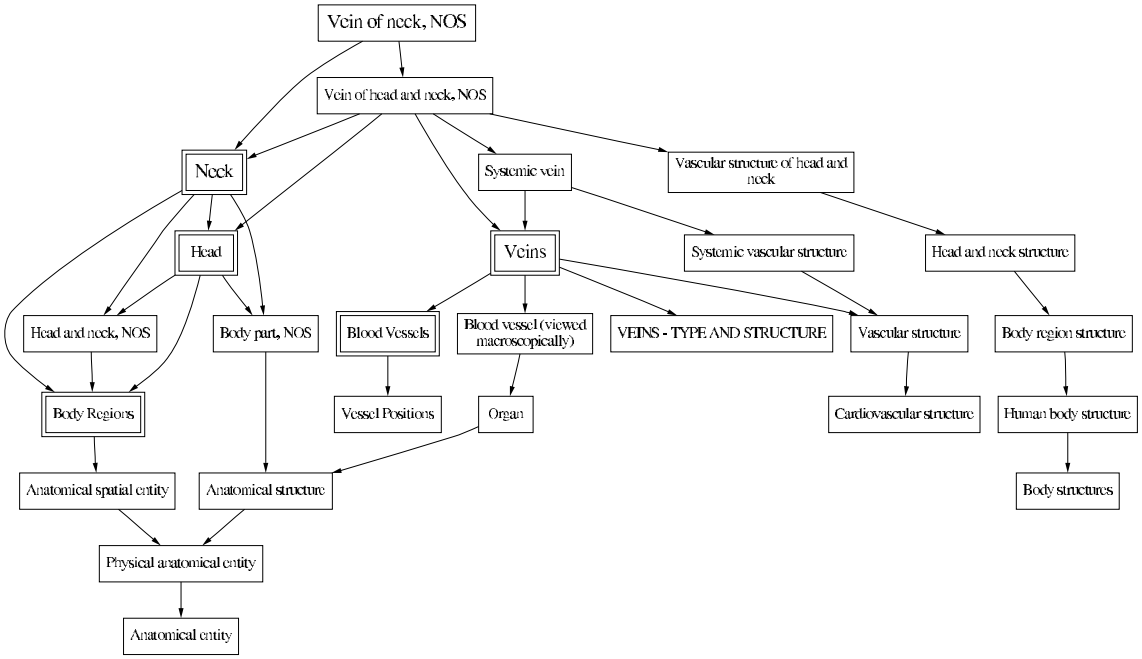

Figure 5. Graph of the ancestors of “Vein of neck, NOS.”
MeSH terms are double framed. The selected MeSH terms are “Neck” and “Veins.” Arrows point to parents or broader concepts.

possible in the ancestor graph, in this case, the only two remaining MeSH terms: “Neck” and “Veins.”

Discussion

This methodology has been evaluated on two sets of 1,000 source concepts each (Bodenreider et al., 1998). Manual review of the results of this test indicated that 61% of the MeSH terms chosen were correct and useful. 28% of MeSH term produced, while not wrong, were too general to be useful, as for example, “Serotonin measurement” mapping to “Laboratory Procedures.” A residue of 11% of the MeSH terms were deemed to be inaccurate. Perhaps not surprisingly, most of the errors were due to step 4 in the algorithm, in which concepts with an unspecified relationship to the source concept are used to generate ancestors.

The algorithm for restricting UMLS concepts to MeSH terms can be tuned from a strict mode (high precision) to a relaxed mode (high recall). Precision can be increased by restricting the seed for the ancestor graph to the source concept itself and not its children or siblings. Further, eliminating step 4 of the algorithm, which appeals to concepts only loosely related to the source concept, also increases precision and reduces the number of inaccurate MeSH terms retrieved. Conversely, making the opposite decision regarding either the seed concept for the ancestor graph or the use of step 4 increases recall. The method that we have described here is an intermediate mode between high precision and high recall, and appears to be optimal in the context of the Indexing Initiative prototype, which ranks and clusters an array of indexing terms based on a range of methodologies.

4.1.5 PubMed Related Citations

The PubMed Related Citations method described here directly computes a ranked list of MeSH headings based on a given title and abstract. It does not use the Restrict to MeSH methodology described in the previous section.

Key Term Extraction

The neighbors of a document (related citations) are those documents in the database that are the most similar to it. The similarity between documents is measured by the words they have in common with some adjustment for document lengths. In order to carry out such a program one must first define what a word is. For us a word is basically an unbroken string of letters and numerals with at least one letter of the alphabet in it. Words end at hyphens, spaces, newlines, and punctuation. A list of 310 common, but uninformative, words (also known as stopwords) are eliminated from processing at this stage. Next a limited amount of stemming of words is done but no thesaurus is used in processing. Words from the abstract of a document are classified as text words. Words from titles are also classified as text words, but words from titles also appear a second time with a special added marker designating them as title words. MeSH terms are placed in a third category and a MeSH term with a subheading qualifier is entered twice, once without the qualifier and once with it. Likewise a MeSH term that is starred (indicating a major concept in a document) is entered once without the star and once with it. These three categories of words (or phrases in the case of MeSH) comprise the representation of a document. No other fields such as author or journal enter into the calculations.

Term Weighting

Having obtained the set of terms that represent each document, the next step is to recognize that not all words are of equal value. Each time a word is used it is assigned a numerical weight. This numerical weight is based on information that the computer can obtain by automatic processing. Automatic processing is important because the number of different terms that have to be assigned weights is close to two million for this system. The weight or value of a term is dependent on three types of information: 1) the number of different documents in the database that contain the term; 2) an estimate of the importance of the term in producing relationships in the database; 3) the number of times the term occurs in a particular document. The first two of these pieces of information are combined to produce a number called the global weight of the term. The global weight is used in weighting the term throughout the database. The third piece of information pertains only to a particular document and is used to produce a number called the local weight of the term in that specific document. When a word occurs in two documents its weight is computed as the product of the global weight times the local weight in each of the documents.

The global weight of a term is greater for the less frequent terms. This is reasonable because the presence of a term that occurred in most of the documents would really tell one very little about any particular document. On the other hand a term that occurred in only one hundred documents out of one million would be very helpful in limiting the set of documents of interest. A word that occurred in only ten documents is likely to be even more informative and will receive an even higher weight. The second factor that enters into the computation of the global weight of a term is what we call the strength of the term. It is defined as the probability that a term that occurs in one document will also occur in any other document that is closely related to the first document. For a

term of a given frequency the higher the strength the greater the global weight. For details of how the global weight is computed for a term we refer the interested reader to Wilbur and Yang (1996), where section 3 is of particular relevance. The local weight of a term within a document is greater the more frequent the term is in that document.

The Similarity Score

The similarity between two documents is computed in two steps. The first step is to add up the weights (local wt1 * local wt2 * global wt) of all the terms the two documents have in common. This provides an indication of how related two documents are. However, this preliminary score suffers from the problem that when a document is scored against a long document and a short document the long document will usually win just because of its length. To correct for this problem we divide this preliminary score by the product of the lengths of the two documents. The resultant score is an example of a vector cosine score. Cosine scoring was originated by Gerard Salton (1988) and has a long history in text retrieval. Our approach differs from other approaches in the way we calculate the local and global weights for the individual terms.

Ranking and Retrieval

Once the similarity score of a document in relation to each of the other documents in the database has been computed, that document's neighbors are identified as the most similar (highest scoring) documents found. These closely related documents are precomputed for each document in MEDLINE so that when you push the button "See Related Articles" the system has only to retrieve this list. This enables a fast response time for such queries.

Neighboring for Indexing

For purposes of indexing one must assume that a document does not have MeSH terms. Hence in our use of neighboring as an indicator of which MeSH terms may be appropriately assigned to a document we perform the neighboring on this document without MeSH terms. This does have the effect of decreasing the accuracy of the neighboring, but the effect is relatively small for documents that possess abstracts. For documents that do not possess abstracts neighboring may not be a useful procedure. However, in such a case the textual analysis approach applied only to titles is also unlikely to be productive of more than a fragmentary and incomplete picture of the desired MeSH representation. Another thing to consider here is the effect over time of changes in the biomedical literature if human indexing were stopped.

4.1.6 The INQUERY Method

A second method of directly computing MeSH headings from input text is the INQUERY method. It is based on the construction of an INQUERY database of records, one for each MeSH heading. INQUERY is a standard statistical, probabilistic, information retrieval system which has the capability of defining multiple indexes on specified text fields. In the method described here, each record consists of several fields containing a variety of information about its MeSH heading:

- KEY: the MeSH heading, itself
- CUI: the concept unique identifier from the Metathesaurus
- SYN: MeSH entry terms plus synonyms from the Metathesaurus
- STY: the concept's semantic type

- MN: MeSH tree numbers
- REL: related concepts, including broader, narrower, and other related concepts from the Metathesaurus
- COT: the top 50 co-occurring Metathesaurus concepts
- PMCIT: the title, abstract and MeSH headings of the top 10 PubMed citations originating from a query using the MeSH heading as Major Topic

The INQUERY method begins by extracting noun phrases from the input text using PhrasEx, the same phrase extraction algorithm used by MetaMap. The extracted phrases are used as query strings to the INQUERY database using a *layered* search strategy enabled by the field indexing mentioned above. The following fields are searched in sequence:

- KEY
- SYN
- REL
- COT
- PMCIT

If INQUERY returns one or more results from one of the above fields, the search process stops, and the INQUERY ranking scores are used to compute mapping scores for the returned records. The final result is the ranked list of MeSH headings corresponding to the ranked records.

4.1.7 Clustering

The ranked lists of MeSH headings produced by all of the methods described so far must be clustered into a single, final list of recommended indexing terms. The task here is to provide a weighting of the confidence or strength of belief in the assignment, and rank the suggested headings appropriately. There are a number of factors that can be recognized as playing a role in that confidence. The method of finding the heading (the path), how much confidence is available in how the method found the heading (the goodness of the match), the location in the text of the nominal phrase that led to that suggestion (the location), and the semantic consistency of the suggested heading with the other suggested headings (the corroborating evidence).

The clustering algorithm uses two formulas for finding the rank score. One is for the weight of a given MeSH heading (or term), the second for the rank order. The term weight formula is

$$TermWeight = TW = \sum_{i=1}^{n} (PathWeight_i \cdot MapScore_i \cdot NavScore_i)$$

where *i* represents a single occurrence of the suggestion of one MeSH heading.

Assigning a weight to the overall method of finding the heading (the Path Weight) allows one to discount a method appropriate to strengths. For example, a certain path might not be very specific, but have some sensitivity in suggesting headings which would otherwise not occur. When headings found by other paths offer corroborative evidence for a heading suggested by this method, the additional confidence gained might be helpful.

The goodness of the match, i.e., how much confidence to place in a given heading, depends on the method used to find the heading. The possibilities are:

- A phrase identified in text is an exact match to a MeSH term. Equivalently, it might have been a match to a UMLS term which was a synonym of a MeSH term.
- Of lesser significance is an exact match to a UMLS term which is then be mapped to a MeSH heading using the Restrict to MeSH method.
- Another possibility is that the phrase is an inexact, or approximate, match to a UMLS term, which is either a synonym of a MeSH heading or mapped to MeSH.

Thus, each time a MeSH heading is suggested, a weighting can be given to that suggestion. This is accomplished using both a *MapScore* and a *NavScore*. The MapScore reflects the confidence in the mapping to a UMLS term, the NavScore the confidence in navigating from a UMLS term to a MeSH Heading.

A ranking score for each suggested MeSH heading can be calculated by the following formula:

$$RankScore = TW \cdot \left[Title + 1 + \sum_{j=1}^{n} (COT_j \cdot TW_j) + \sum_{k=1}^{n} (REL \cdot TW_k) \right]$$

where j and k represent other suggested MeSH headings, semantically related to the suggested heading by either co-occurrences in MEDLINE or by occurrence in the same hierarchy in MeSH.

With regard to the importance of location, the main consideration was whether or not the phrase leading to a heading suggestion was mentioned in the title. All other things being equal, indexers know that things mentioned in the title of the article are probably more important than other concepts mentioned in the article. Similarly, if the heading was suggested by a phrase occurring in the title, it should be given more weight. The additional weight is added as a constant in the formula.

Semantic consistency can be thought of as corroborative evidence for the goodness of a suggestion. It is identified by relationships that a suggested heading has with other suggested headings. These relationships might be either the occurrence in the same hierarchy (as parents or siblings), or as known co-occurring headings in MEDLINE. This latter evidence needs to be weighted according to a normalized frequency of this co-occurrence. The normalized frequency times a constant becomes the COT weight. The former evidence is the REL weight, and is a simple constant.

The overall Rank Score can be altered by changing any of the constants (COT, REL, Title, PathWeight) or by changing the method by which the weight is calculated (NavScore and MapScore). Altering these values allows a number of experiments to be performed to evaluate the robustness of the weighting scheme, and to establish reasonable values for the constants.

4.2 Experiments

Many experiments were conducted using a randomly selected sample of 200 MEDLINE citations with entry month of January 1998. Each experiment consisted of processing the citations with a given set of prototype parameter values. Of particular interest were the weights assigned to the alternative paths for concept discovery. Other parameters, such as the navigational and map score,

were set to their default values. Table 1 shows the results of the experiments for single paths, i.e.,

MMI	Tri	Bar	Rel	Inq	5 P	5 R	10 P	10 R	20 P	20 R	40 P	40 R
1	0	0	0	0	.55	.26	.38	.33	.25	.37	.19	.39
0	1	0	0	0	.50	.25	.37	.32	.23	.38	.15	.41
0	0	1	0	0	.46	.21	.34	.28	.23	.33	.17	.35
0	0	0	1	0	.45	.17	.33	.20	.25	.34	.20	.39
0	0	0	0	1	.25	.10	.18	.14	.13	.20	.09	.27

Table 1. Single-path experiments

where a given path was given a weight of 1 and all other paths a weight of 0. In the table, MMI, Tri, Bar, Rel and Inq denote the paths for the MetaMap Indexing, Trigrams, Barrier Words with Approximate Matching, PubMed Related Citations and INQUERY methods, respectively. For each experiment, the recommended indexing terms were compared with the terms assigned by NLM indexers. Precision (P) and recall (R) values are computed for the top 5, 10, 20 and 40 recommended indexing terms. The experiments show that the MetaMap Indexing path is the single strongest path and that the Trigram and PubMed Related Citations paths perform well as more recommended terms are considered.

A second series of experiments involving more than one path is shown in Table 2. The best results

MMI	Tri	Bar	Rel	Inq	5 P	5 R	10 P	10 R	20 P	20 R	40 P	40 R
1	1	0	0	0	.51	.25	.38	.33	.24	.39	.15	.42
1	0	1	0	0	.49	.24	.37	.32	.24	.38	.17	.42
1	0	0	1	0	.57	.27	.42	.36	.30	.48	.20	.58
1	0	0	0	1	.43	.21	.34	.29	.23	.38	.14	.44
0	1	1	0	0	.50	.25	.38	.33	.25	.40	.15	.44
0	1	0	1	0	.56	.28	.44	.38	.30	.49	.19	.57
0	1	0	0	1	.50	.25	.37	.33	.24	.40	.15	.46
0	0	1	1	0	.55	.25	.42	.35	.29	.46	.19	.55
0	0	1	0	1	.46	.22	.35	.30	.23	.36	.14	.42
0	0	0	1	1	.49	.20	.36	.28	.25	.39	.16	.48
24	0	0	1	0	.60	.29	.48	.41	.32	.51	.20	.61
96	1	0	4	0	.58	.29	.47	.41	.31	.51	.20	.62
96	2	0	3	0	.59	.30	.47	.41	.32	.51	.20	.61

Table 2. Multi-path experiments

for two paths with equal weights are obtained with the MetaMap Indexing and Related Citations paths. The most striking result, however, is that adding the Related Citations path to another path always improves results. Further experiments applying unequal weights to two paths showed that the best weighting of two paths occurred with a 24:1 ratio of MetaMap Indexing to Related Citations. Finally, experiments giving weight to the Trigrams path in addition to the MetaMap Indexing and Related Citations never produced an overall result better than the best weighting for the

two paths alone although some weightings performed better for larger numbers of recommended terms.

Another set of experiments consisted of combining elements from some of the original paths focusing on the method of determining noun phrases and the concept discovery methods for these phrases. One of four noun phrase methods was combined with one of two concept discovery methods. The noun phrase methods are:

- PhrasEx from the MetaMap Indexing and INQUERY paths;
- the Barrier Word method;
- Chopper, based on the MIT parser; and
- Frazer, a modification of PhrasEx.

The two concept discovery methods are:

- Semantic mode from the MetaMap Indexing path; and
- MetaMap's Browse mode, the precursor of Approximate Matching.

None of the combined phrase extraction and concept discovery methods performed better than the original combination forming the MetaMap Indexing path.

As a result of the experiments described above, the IND Prototype was created using the 24:1 ratio of weighting the MetaMap Indexing and Related Citations paths.

4.3 An Example

We now give an example of the automatic indexing produced by the current IND Prototype. Consider the following MEDLINE citation:

> UI - 98018928
>
> TI - Bupivacaine inhibition of L-type calcium current in ventricular cardiomyocytes of hamster.
>
> AB - BACKGROUND: The local anesthetic bupivacaine is cardiotoxic when accidentally injected into the circulation. Such cardiotoxicity might involve an inhibition of cardiac L-type Ca2+ current (ICa,L). This study was designed to define the mechanism of bupivacaine inhibition of ICa,L. METHODS: Cardiomyocytes were enzymatically dispersed from hamster ventricles. Certain voltage- and time-dependencies of ICa,L were recorded using the whole-cell patch clamp method in the presence and absence of different concentrations of bupivacaine. RESULTS: Bupivacaine, in a concentration-dependent manner (10-300 microM), tonically inhibited the peak amplitude of ICa,L. The inhibition was characterized by an increase in the time of recovery from inactivation and a negative-voltage shift of the steady-state inactivation curve. The inhibition was shown to be voltage-dependent, and the peak amplitude of ICa,L could not be restored to control levels by a wash from bupivacaine. CONCLUSIONS: The inhibition of ICa,L appears, in part, to result from bupivacaine predisposing L-type Ca channels to the inactivated state. Data from washout suggest that there may be two mechanisms of inhibition at work.

Bupivacaine may bind with low affinity to the Ca channel and also affect an unidentified metabolic component that modulates Ca channel function.

MH - Anesthetics, Local/*PHARMACOLOGY
Animal
Bupivacaine/*PHARMACOLOGY
Calcium Channels/*DRUG EFFECTS
Dose-Response Relationship, Drug
Hamsters
Heart/*DRUG EFFECTS
Male
Support, Non-U.S. Gov't

The human indexing for this citation has nine MeSH headings four of which are check tags. The Prototype produces 125 recommended MeSH headings some of which are listed in Table 3 in

N	MeSH Heading	Rank Score	MMI	Rel Cit
1	**Calcium Channels**	86802	x	x
2	Calcium	26581	x	x
3	**Bupivacaine**	23809	x	x
4	Calcium Channel Blockers	23103		x
5	Membrane Potentials	21353		x
6	Myocardium	15906		x
7	**Anesthetics, Local**	13671	x	
8	**Heart**	8976	x	x
9	Heart Ventricle	8350	x	x
10	Potassium Channels	6665		x
11	Patch-Clamp Techniques	6495		x
12	Ryanodine	6492		x
13	Dihydropyridines	4864		x
14	Egtazic Acid	4860		x
15	Myocardial Contraction	4377		x
	...			
51	Anesthetics, Intravenous	478		x
52	Time	419	x	
53	**Dose-Response Relationship, Drug**	399		x
54	Receptors, Adrenergic, beta-1	364		x
55	Cyclosporine	355		x
	...			

Table 3. Prototype recommended indexing (excluding check tags)

ranked order. The Prototype finds all five headings which are not check tags; these are shown in bold in the table. Note, however, that the rank score for "Dose-Response Relationship, Drug" is very low. Also, the Prototype finds the check tags "Animal" and "Hamsters" but not the check tags "Male" and "Support, Non-U.S. Gov't".

It is interesting to note that while both the MetaMap Indexing (MMI in the table) and PubMed Related Citations (Rel Cit) methods found most of the headings, only MMI found “Anesthetics, Local” and only Rel Cit found “Dose-Response Relationship, Drug”.

Further analysis of the results shows that the Prototype produced additional useful indexing terms:

- “Calcium”: The “Calcium Channels’” discussion in the citation includes reference to the movement of calcium ions across cell membranes; so “Calcium/METABOLISM” is a possible heading/subheading combination;
- “Heart Ventricle”: The cardiomyocytes are taken from the heart ventricle;
- “Calcium Channel Blockers”: In both the title and abstract, it is clearly stated that bupivacaine has the action of calcium channel inhibition;
- “Membrane Potentials”: This heading is appropriate for indexing because voltage and voltage shift are discussed in the abstract; and
- “Patch-Clamp Techniques”: This method is also described in the abstract.

4.4 Journal Descriptor (JD) Indexing

Background

As part of the research underlying the Indexing Initiative, we are investigating a novel approach to fully-automated indexing (Humphrey 1998; Humphrey 1999) based on NLM’s practice of maintaining a subject index to journal titles using terms corresponding to specialties associated with biomedicine. This journal descriptor (JD) indexing is meant to complement the methods described earlier in this report.

JD’s are a set of 141 preferred MeSH terms designated for use in indexing MEDLINE journals by subject. Of the 4,000 MEDLINE serials, 3,330 are assigned JD’s; of these, about 74% are assigned a single JD, 21% just two JD’s, and the remainder up to five JD’s. For example, the *American Journal of Cardiology* has the JD Cardiology.

Since all citations inherit the JD’s of their respective journals, JD’s can be thought of as indexing terms for documents as well as for journals, and, in fact, JD’s have often been used indirectly by professional searchers. For example, to retrieve literature on neurotransmitters in the field of cardiology, the search term Neurotransmitters may be intersected with the JD Cardiology, which can only be searched by specifying the title abbreviations or codes for journals with this JD.

There are several impediments to using JD’s as currently implemented for MEDLINE retrieval. The considerable difficulty in accessing JD’s in available retrieval systems and the fact that some journals have no JD seem not to be research issues. JD indexing research addresses the problem that retrieval based on searching a JD as an inherited descriptor is restricted to the particular journals having this JD. For example, a cardiology citation in the New England Journal of Medicine inherits only the JD assigned to this journal, which is Medicine, and thus cannot be retrieved by the JD Cardiology. The research discussed here investigates the feasibility of automatically assigning JD’s to all MEDLINE citations, as appropriate, regardless of the JD’s currently assigned.

Relying on the intellectual effort of NLM catalogers who maintain current information about NLM serials, JD indexing has the potential to supplement other forms of automatic indexing by

providing powerful access points in certain types of searching, much as is done by search engines on the Web that organize documents in very general categories, thereby subdividing the Web into domain-specific information sources. Further applications being investigated include using JD indexing as an aid to resolving word-sense ambiguity in natural language processing.

Methodology

The methodology for using JD's as indexing terms is based on the characterization of a word by a "JD profile." The JD profile associates the JD's of journals with words commonly occurring in titles and abstracts of papers in these journals and is computed from a training set of MEDLINE citations to these papers. Once the JD profile has been computed for each word occurring in the training set, it can be used as the basis for indexing any medical text outside the training set, including but not limited to MEDLINE citations. This automatic indexing relies on a calculation based on a composite of the JD profiles for the words occurring in the document to be indexed. The ranked list of JD's resulting from this calculation become the indexing terms for the document under consideration.

The association captured in the JD profile is based either on the number of occurrences of the word in the training set or on the number of citations in the training set containing the word. In the former case, the number of occurrences of a word in association with a particular JD is divided by the total number of occurrences of this word, while in the latter, the number of citations associated with a particular JD and containing this word is divided by the total number of citations containing this word. Both methods of computing the JD profile are under consideration. For brevity, only one method, based on citation count, will be illustrated in this written report.

For example, the word JD profile for *chemotherapy* is as follows, using our current training set of 21,760 citations. *Chemotherapy* occurs a total of 657 times in 304 citations, and forty-eight JD's have been associated with these citations by NLM. The distribution of *chemotherapy* in association with the JD's occurring in the training set is listed in Figure 6 as a percentage of the total number of citations in which this word appears (only the most frequent associations are given). For instance, the frequency of occurrence of *chemotherapy* in association with the top-ranked JD Medical Oncology (0.453947) is determined by dividing the number of citations associated with Medical Oncology and containing *chemotherapy* (138) by the total number of citations in which this word appears (304).

CITATION COUNT FOR WORD PER JD / TOTAL CITATION COUNT, BY COUNT:

|Medical Oncology| 138/304 = 0.453947

|Hematology| 39/304 = 0.128289

|Medicine| 31/304 = 0.101974

|Pediatrics| 24/304 = 0.078947

|Surgery| 16/304 = 0.052632

etc.

Figure 6. The JD profile for *chemotherapy*

The principle underlying the feasibility of JD's as document descriptors is the fact that the distribution of words in citations associated with a particular JD is not uniform. Frequency of occurrence is purported to correlate with the semantic content of the text of the citation. In this example, *chemotherapy* occurs most frequently associated with the JD Medical Oncology, considerably less so with Hematology, and so on. The association of a word with a particular array of JD's forms the basis for automatic indexing. For example, once the JD profile for *chemotherapy* has been computed from the training set, this profile functions as an indicator of the semantic content of any text in which *chemotherapy* occurs. JD's derived from the JD profiles for all the words in a document can then be used as indexing terms for that document.

To illustrate indexing a document outside the training set, we can use the following title from the *New England Journal of Medicine*: "Dexamethasone, Granisetron, or Both for the Prevention of Nausea and Vomiting during Chemotherapy for Cancer." Considering this title as a document, the top-ranked JD's assigned as indexing terms are shown in Figure 7. (Note that although this journal has the JD Medicine assigned by NLM, this assignment is not used in the JD indexing of this text.)

JD'S AND RANK BASED ON CITATION COUNT FOR WORD, BY RANK:
|Medical Oncology| 0.18495
|Medicine| 0.105122
|Pharmacology| 0.00679
etc.

Figure 7. The top-ranked JD's for "Dexamethasone, Granisetron, or Both for the Prevention of Nausea and Vomiting during Chemotherapy for Cancer."

In order to arrive at these terms, tables like the following, associating words in the text with the JD Medical Oncology are computed for each JD. Table 4 shows how many times words from the document occur associated with the JD Medical Oncology in the training set based on citation count. For example, *granisetron* occurs in six citations in the training set and three of these have the JD Medical Oncology (i.e., are from Medical Oncology journals). As noted earlier in the JD profile for *chemotherapy* this word occurs in 304 citations and 138 of these are associated with Medical Oncology. The ranking (0.18495) of this JD as an indexing term for the text under consideration is computed by averaging the percentages given in the third column of the table. The fact that Medical Oncology was the top-ranked JD in the JD profiles for five words in the text (*chemotherapy,* as illustrated earlier, as well as *granisetron*, *nausea*, *vomiting*, and *cancer*) contributed to this being the top-ranked JD for the text.

After experimenting with using *ad hoc* stopword lists and word variants it was decided to use a more general method of considering all words in the text (with more than two characters) and eventually using statistical weighting to compensate for raw distributions which do not reflect meaning.

Training Set

The current training set consists of 21,760 citations from the July, 1995, MEDLINE file having a 1995 publication date. The distribution of JD's assigned to citations in this set is given in Table 5

WORD IN DOCUMENT TO BE INDEXED	Medical Oncology CITE COUNT/ TOTAL CITE COUNT	Medical Oncology RANK (CITE COUNT)
DEXAMETHASONE	1/68	0.014706
GRANISETRON	3/6	0.5
BOTH	147/4038	0.036404
FOR	570/11837	0.048154
THE	740/17625	0.041986
PREVENTION	9/317	0.028391
NAUSEA	19/51	0.372549
AND	736/17301	0.042541
VOMITING	13/45	0.288889
DURING	119/3050	0.039016
CHEMOTHERAPY	138/304	0.453947
CANCER	320/907	0.352811

Table 4. Words associated with the JD Medical Oncology

for the ten most frequently-assigned JD's. (Citations from journals not assigned a JD (1% of the total) were eliminated from the training set).

NO. OF CITATIONS	JD	% OF CITATIONS
2,467	Medicine	11%
1,585	Biochemistry	7%
972	Nursing	4%
851	Medical Oncology	4%
828	Surgery	4%
768	Allergy and Immunology	4%
681	Pharmacology	3%
632	Science	3%
584	Neurology	3%
576	Biotechnology	3%

Table 5. Most frequently assigned JD's in the training set

Biases presumably not representative of MEDLINE are a consequence of the relatively small size of this training set. For example, the citation count for Nursing is abnormally high simply because an unusually large number of issues of journals in that discipline were indexed during July, 1995. Such biases degrade the results of indexing with this method. The construction of a considerably larger training set is being pursued in order to address this problem.

Any training set representative of MEDLINE, even a large one, will reflect the inherent biases of MEDLINE with regard to discipline. Citation counts for a discipline reflect not only the number

of journals indexed in a discipline, but also frequency of publication and number of articles per issue of such journals. Word counts for a discipline reflect in addition the length of titles and number and length of abstracts keyed in per indexed journal. Again, once a larger training set is constructed, a truer picture of these inherent biases should emerge so that techniques to deal with them can be explored.

Current research

The thrust of current research underlying this project is to improve system performance. Ways of doing this include enlarging the training set to better represent domain distributions in MEDLINE, trying standard statistical Information Retrieval methods such as term weighting, developing statistical and other methods to compensate for under- and over-representation of certain domains, investigating problems in JD assignments, and exploring methods that associate JD's with elements of MEDLINE citations other than individual words.

JD's assigned to journals that are not representative of every paper in the journal are a particular problem. A solution may be to combine certain JD's, for example, Medical Oncology and Neoplasms, Experimental combined as Neoplasms. Associating JD's with elements other than individual words presumes the availability of techniques for identifying these elements. An example of such an element would be automatically-generated noun phrases.

Since the system does not use the JD assignment for a test article, a ready-made measure of system performance can be whether the system recommends this JD, for example, whether the system assigns the JD Cardiology to test documents from the *American Journal of Cardiology.* An evaluation may also be based on any of the test collections available generally in the Indexing Initiative, including the recently created CBM (*Current Bibliographies in Medicine*) collection.

5. Evaluation

5.1 Background

Evaluation constitutes an integral part of the research supporting the development of automated methods for assigning indexing terms to MEDLINE abstracts. The methodology being pursued adheres to standard practice in information retrieval (IR) research (Cleverdon, Mills, and Keen, 1966; Sparck Jones, 1981; Tague-Sutcliffe, 1992). The ultimate goal of any IR system is user satisfaction. However, the complex interaction of the many constituent components in such a system makes it challenging to assess precisely the effect of any one of these components on overall success (Soergel, 1994). Therefore, multiple types of evaluation (Saracevic, 1995) are required in order to determine the likely effect of the changes being pursued in the Indexing Initiative.

IR systems can be evaluated at several levels, including those concerned with effectiveness of the underlying hardware and software, input and output procedures, and overall user satisfaction. Saracevic (1995) notes that assessment of these levels naturally falls into two evaluation categories: system-centered and user-centered; he further comments that these approaches are complementary and that both are ultimately required for effective evaluation of IR systems. Currently, we are particularly concerned with that part of the retrieval algorithm dependent on the indexing terms assigned to documents and are thus pursuing a system approach to evaluation. Ultimately, a more user-oriented strategy will be required to complement these efforts.

Evaluation techniques indicate how the Indexing Initiative Prototype compares to current indexing practice and also provide a guide for improvements. Human-assigned MeSH terms serve as a *de facto* standard, and our research seeks to at least achieve current levels of effectiveness with automatic means. In order to compare the automatic methods to human indexing, we have initiated techniques centered around two broad strategies: index-based evaluation and retrieval-based evaluation. In the index-based approach we compare the specific indexing terms suggested by automatic methods to those assigned by humans for a particular abstract. The retrieval-based method compares the effectiveness of automatically-generated terms against human-assigned terms in the context of a test collection of queries and relevant documents. In both types of evaluation we employ the standard IR evaluation metrics of recall, precision, and average precision.

5.2 Index-based Evaluation

Index-centered evaluation is conceptually straightforward and is relatively easy to implement. For each abstract under consideration the automatically-generated terms are compared to the MeSH terms assigned by humans. The central weakness underlying such evaluation is the assumption that the MeSH terms assigned by humans are uniquely optimal for representing the content of the relevant document. A set of terms other than the human-assigned MeSH terms may be equally effective with respect to retrieval. Nevertheless, current MeSH indexing constitutes a known standard against which to judge progress in the Indexing Initiative. In assessing the results of the current prototype against this standard, the automatically-generated terms were compared by exact concept match to the human-assigned terms for the 200 MEDLINE citations used for testing the Indexing Prototype. The 11-point average precision for this comparison was 34%.

Error analysis of these results indicates that the Indexing Initiative deviations from the standard largely fall into three categories: a) false positives due to word sense ambiguity, b) false negatives that could only be found in the full-text article, and c) closely related terms that did not satisfy the criterion of an exact concept match.

False positives, that is, automatically-assigned terms that are not in the set of terms assigned by humans (and which do not accurately reflect the content of the relevant document) are almost always due to word sense ambiguity. This happens, for example, when the Prototype suggests the indexing term “Inhibition (Psychology)” based on the text *Bupivacaine inhibition of L-type calcium*. False negatives, that is, terms which were assigned by the human indexers but which were not assigned by the automatic method often occur because the relevant concept is not mentioned in the abstract, but rather in the full text of the article on which the human indexers base their analysis. Further research is being planned to correct both types of errors.

Some Indexing Initiative terms that do not exactly match the humanly assigned terms are nonetheless semantically close, and research is being pursued to address this phenomenon. Effective indexing is ultimately based on representation of the semantic content of a document. If alternative sets of indexing terms can adequately represent the meaning of a particular document, such sets, while differing in detail, would occupy the same “semantic space.” Recent research in the Indexing Initiative (Bodenreider and McCray, 1999) based on the notion of semantic locality (Nelson et al., 1991) seeks to determine semantic relatedness between biomedical concepts. The methodology being pursued computes a semantic proximity score for any given pair of UMLS concepts using hierarchical and non-hierarchical relationships as well as the co-occurrence of concepts in the biomedical literature. Individual proximity scores can then be aggregated in order

to compare sets of concepts. An application of the method will be used to calculate the semantic distance between the set of MeSH descriptors suggested by the Indexing Initiative Prototype for a given abstract and those assigned by human indexers. This “semantic proximity index” may more accurately reflect the effectiveness of the automatically-generated indexing terms in comparison to the human-assigned terms than is indicated by an exact concept match between the two sets of terms.

An experiment complementary to the exact concept matching further assessed the computer-generated terms in comparison to those assigned by the human indexers. In this study, several experienced indexers were asked to evaluate automatically-assigned indexing terms during the process of indexing particular journal articles. The indexers completed surveys to record their comments on the suggested terms. In addition to false negatives and positives of the types mentioned above, the indexers noted a number of automatically generated terms which were too general to be useful as document descriptors.

5.3 Retrieval-based Evaluation

Retrieval-centered evaluation is traditional in the IR field (Salton, 1992) and is reasonably well understood. Further, the results are not dependent on specific indexing terms as is the case with index-based evaluation. However, a test collection, with relevance judgments, is needed.

Traditionally test collections contain tens of thousands of documents, but there is a trend for recently created test collections to contain hundreds of thousands of documents (Harman, 1996). A potential confound in the use of a test collection is that bias in the collection may skew results (Korfhage & Yang, 1991). A further concern is that the relevance judgments in the collection may not reflect the assessments of actual users (Schamber, 1994). Employing a large test collection requires substantial computing resources.

We have several resources available to address the potential problems associated with using a test collection in retrieval-based evaluation. In order to mitigate the effects of bias in any one collection we plan to evaluate the Indexing Initiative Prototype against three small (Schuyler, McCray, and Schoolman, 1989; Hersh, Hickam, Haynes, and McKibbon, 1994; Wilbur, 1996) and two large (Hersh, Buckley, Leone, & Hickam, 1994; Bean, et al., 1999) test collections. All five collections consist of queries with associated relevant MEDLINE citations. The three small collections contain roughly 3,000 documents each, while the large ones consist of more than 300,000 citations each.

A new large biomedical test collection was recently created as part of Indexing Initiative research (Bean, et al., 1999). An innovative aspect of this work is that the relevance judgments are based on assessments of actual users. The CBM Test Collection was constructed on the basis of six recent issues in NLM’s *Current Bibliographies in Medicine* (CBM) series. Each bibliography in this series covers a distinct subject in biomedicine and was prepared in support of a specific NIH Consensus Development Conference to address a set of questions which represent the statement of information needs targeted by the CBM. Relevant citations for each question were identified by a team consisting minimally of a professional reference librarian and a subject expert, using reiteratively refined search strategies and topical reference lists solicited from other subject experts, and organized topically. Thirty-three baseline natural language queries were back-generated by precise matching of questions with headings and subheadings in the classification. Citations in

each section were mapped to a test file of 388,900 1995 MEDLINE documents, yielding 1,266 query document pairs.

In order to accommodate the large amount of text which must be processed for retrieval-centered evaluation of the Indexing Initiative Prototype, software has recently been devised based on a parallel, distributed model which allows 1,000 MEDLINE citations to be processed per hour. This resource accommodates the hundreds of thousands of documents occurring in the large test collections to be processed in a reasonable amount of time, and evaluation of the current Prototype is about to begin on these collections.

5.4 User-centered Evaluation

The ultimate goal of any IR system is user satisfaction, regardless of the underlying technology. Such satisfaction is determined by numerous factors beyond the technical ability of a system to deliver topically relevant documents. The conclusion reached by many investigators is that a more user-oriented notion of retrieval system evaluation is needed in order to address these issues (Harman, 1992; Su, 1992; and Gluck, 1996), and recent system development in IR is often assessed with the user in mind (Jose, Furner, and Harper, 1998, for example).

Discussions are currently underway in the Indexing Initiative considering possible approaches to the design of a user-oriented evaluation study. Several recent studies serve as a guide in this regard. Hersh, Pentecost, and Hickam (1996) report on an interesting, task-oriented evaluation strategy in a biomedical setting, which focuses on the user's information need. Methodologies are being developed in the context of the TREC experiments (Beaulieu, Robertson, and Rasmussen 1996) which provide a means of accommodating the user in formal IR experiments. Surveys of the type reported in Lindberg, et al. 1993b can provide valuable insight into the impact that an IR system has on the professional activities of users.

6. Project Plan

Continued research supporting automatic indexing first concentrates on technical development of the current Prototype and then seeks to increase the indexing effectiveness of the automatic indexing process. Efforts toward enhancement are driven by the error analysis conducted during evaluation and by the Journal Descriptor research discussed above.

6.1 Technical Development of the Prototype

In order to be able to perform large numbers of experiments with the Indexing Initiative Prototype without incurring significant computational delays, we used a fixed set of 200 MEDLINE citations precomputing the recommended indexing terms for each of the five basic methods. Now that the best combination of methods has been determined and embodied in the Prototype, the process needs to be made dynamic for more realistic testing. Such a technical development effort is the first step toward enabling the enhancements described below.

6.2 Enhancements to the Prototype Based on Results of Evaluation

Initial results of error analysis during the evaluation process indicate two major areas of focus for continued enhancement of the Indexing Initiative Prototype. The first of these addresses the problem of word sense ambiguity, a phenomenon of natural language which presents a particular challenge to all methods concerned with automatically processing free text. As noted above in the section on evaluation, errors due to word sense ambiguity decrease precision when comparing automatically generated indexing terms against those assigned by humans. The second major area of planned research confronts the fact that, currently, the automatic methods of assigning indexing terms apply only to the title and abstract for a particular journal article, while human indexers base their analysis on the full text of the article. This restriction causes the computer-generated terms to suffer recall errors in comparison to the humanly assigned document descriptors.

6.2.1 Word Sense Disambiguation

Indexing errors due to word sense ambiguity arise when the UMLS Metathesaurus has a single string referring to two or more distinct concepts. Such cases are disambiguated in the Metathesaurus; however, with current processing, the Indexing Initiative Prototype does not have a means of choosing which concept is appropriate in the given textual context. Current research in statistically-based natural language processing (Manning and Schutze, 1999) addresses automatic resolution of this type of ambiguity. One challenge in this method is that it requires a significant amount of training text, which must often be disambiguated by hand. We have initiated research in a memory-based learning approach (Daelemans, 1995) which minimizes this effort by first concentrating on non-ambiguous training text. In addition the work on Journal Descriptors described in Section 4.4 offers another promising approach to word sense disambiguation.

6.2.2 Full Text Processing

Two approaches are planned to meet the challenge of basing automatic indexing on the full text of the article being processed. The first of these involves submitting the full text of journal articles to the automatic indexing process. Optimal results are likely to be achieved by addressing those sections of a full-text article which concentrate on the main points of the article. Considerable research in the field of computational linguistics (Lin and Hovy, 1997, for example) is concerned with identifying key topics and sections in a full-text article. Insights from human indexer practice in this regard provides guidance for the automatic methods being developed. For example, in a preliminary study on the effect of key sentences on MetaMap Indexing results, we used the observation of an expert indexer that the last (and sometimes the first) sentence of the introduction of a full journal article often supplies crucial information about how to index the article.

6.2.3 Semantic Proximity

A second method addressing the fact that automatic indexing is based on the abstract of an article, rather than the full text, seeks to increase the number of indexing terms for a specific article by appealing to the notion of semantic locality (Nelson et al., 1991). Recent research (Bodenreider and McCray, 1999) based on this notion provides a precise way of computing the “semantic distance” between a given pair of UMLS concepts (either one of which may be an indexing term for a specific article). This “semantic proximity score” is based on hierarchical and non-hierarchical

relationships between terms in the UMLS as well as the co-occurrence of concepts in the biomedical literature. Research is planned which investigates the use of semantic proximity to suggest additional indexing terms based on specific configurations in the semantic space defined by an existing set of automatically-generated terms.

6.3 Enhancements Based on Journal Descriptor Indexing

The research described earlier on indexing using JD's can be used in a number of ways in the Indexing Initiative. Each of these ways depends on its ability to associate JD's with a word, a phrase or a large body of text. Beyond the ability to explicitly limit a given search to documents associated with given JD's, it can also be useful for performing some disambiguation tasks such as when a word occurs more frequently in one domain than another. Using it to determine the *domain* of a user's query can significantly improve retrieval performance by searching only in the appropriate domain.

6.4 Enhancements Based on Machine Learning

The document neighboring algorithm for finding related citations in PubMed makes use of the prior human MeSH assignments to documents already in the database to guide the assignment of MeSH terms to a new document. Those documents that achieve a high similarity score in relation to a new document receive the most influence in determining the assignment of MeSH terms to the new document. This is only one way in which the data on previously indexed documents can be used to make an assignment of MeSH terms to a new document. Another approach is to make overt use of the various machine learning techniques that have been developed in recent years to learn the context of previously assigned MeSH terms. Once such a context has been learned one then only needs to test a new document to see if it qualifies as a context for the MeSH term under consideration. The strength of assignments is generally a numeric score and those MeSH terms with the highest scores may be assigned to the document. Methods that seem particularly interesting in this regard are naïve Bayes (Elkans, 1997; Langley, Iba, & Thompson, 1992; Langley & Sage, 1994), adaptive boosting (Freund & Schapire, 1995), and support vector machines (Platt, 1998; Platt, 1999). The plan is to compare these methods with the current PubMed neighboring approach.

7. Summary

The Indexing Initiative began with the realization that the quantity of biomedical literature is growing dramatically in the context of limited resources (especially experienced indexers) available for indexing that literature. Early IND efforts consisted of a disparate collection of research projects examining various aspects of the indexing problem. Recent work has focused the results from the initial projects on the development of the Indexing Prototype, a system for testing alternative indexing strategies. The plan described in the previous section will guide future efforts to make the transition from Indexing Prototype to deployable system. Thus our near- and mid-term efforts will concentrate on technical and substantive enhancements to the Prototype together with the concomitant evaluation exercises throughout the process.

8. References

Aronson, A. R. (1996). The effect of textual variation on concept based information retrieval. *Proceedings of AMIA Annual Fall Symposium*, 373-7.

Aronson, A. R., & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. *Proceedings of AMIA Annual Fall Symposium*, 485-9.

Aronson, A. R., Rindflesch, T. C., & Browne, A. C. (1994). Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94*, 197-216.

Bean, C. A., Selden, C. R., Aronson, A. R., & Rindflesch, T. C. (1999). From bibliography to test collection: Enhancing topical relevance assessment for bibliographic information retrieval system evaluation. *Proceedings of AMIA Annual Fall Symposium*, (to appear).

Beaulieu, M., Robertson, S., & Rasmussen, E. (1996). Evaluating interactive systems in TREC. *Journal of the American Society For Information Science, 47*(1), 85-94.

Bodenreider, O., & McCray, A. T. (1999). Towards a semantic proximity score between biomedical concepts, (unpublished).

Bodenreider, O., Nelson, S. J., Hole, W. T., & Chang, H. F. (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proceedings of AMIA Annual Fall Symposium*, 815-9.

Cimino, J. J., Johnson, S. B., Peng, P., & Aguirre, A. (1993). From ICD9-CM to MeSH using the UMLS: a how-to guide. *Proceedings of Annual Symposium on Computer Applications in Medical Care (SCAMC)*, 730-4.

Cleverdon, C. W., Mills, J., Keen, E. M., & Cranfield Research Project. (1966). *Factors determining the performance of indexing systems (Volume 1: Design; Volume 2: Test results)*. Cranfield (Beds.),: College of Aeronautics.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A practical part-of-speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*.

Daelemans, W. (1995). Memory-based lexical acquisition and processing. In P. Steffens (Ed.), *Machine translation and the lexicon: third International EAMT Workshop, Heidelberg, Germany, April 26-28, 1993: proceedings* (pp. 85-98). Berlin; New York: Springer-Verlag.

Elkans, C. (1997). *Boosting and naive Bayesian learning* (Technical Report CS97-557). La Jolla, California: University of California, San Diego.

Freund, Y., & Schapire, R. E. (1995). *A decision-theoretic generalization of on-line learning and an application to boosting.* (Unpublished manuscript available from authors over the web). Murray Hill, New Jersey: AT & T Research.

Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. *Information Processing & Management, 32*(1), 89-104.

Harman, D. (1992). Evaluation Issues in Information-Retrieval. *Information Processing & Management, 28*(4), 439-440.

Harman, D. (1996). Panel: Building and using test collections. In H.-P. Frei (Ed.), *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 335-337).

Hersh, W. R., Buckley, C., Leone, T. J., & Hickam, D. H. (1994a). OHSUMED: An interactive retrieval evaluation and new large scale test collection. In W. B. Croft & C. J. Rijsbergen (Eds.), *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 192-201).

Hersh, W. R., Hickam, D. H., Haynes, R. B., & McKibbon, K. A. (1994b). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *J Am Med Inform Assoc, 1*(1), 51-60.

Hersh, W. R., Pentecost, J., & Hickam, D. H. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society For Information Science, 47*(1), 50-56.

Humphrey, S. M. (1998). A new approach to automatic indexing using journal descriptors. *Proceedings of the ASIS Annual Meeting, 35*, 496-500.

Humphrey, S. M. (1999). Automatic indexing of documents from journal descriptors: A preliminary investigation. *Journal of the American Society For Information Science, 50*(8), 661-674.

Jose, J. M., Furner, J., & Harper, D. J. (1998). Spatial querying for image retrieval: a user-oriented evaluation. In W. B. Croft (Ed.), *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 232-240).

Korfhage, R. R., & Yang, J. J. (1991). A cautionary tale. *SIGIR Forum, 25*(2), 104-5.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers, *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 223-228). San Jose: AAAI Press.

Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers, *Proceedings of the Tenth Conference on Uncertainty in artificial intelligence* (pp. 399-406). Seattle, WA: Morgan Kaufmann.

Lin, C., & Hovy, E. (1997). Identifying topics by position. *Proceedings of the Fifth Conference on Applied Natural Language Processing (Association for Computational Linguistics)*, 283-290.

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993a). The Unified Medical Language System. *Methods Inf Med, 32*(4), 281-91.

Lindberg, D. A., Siegel, E. R., Rapp, B. A., Wallingford, K. T., & Wilson, S. R. (1993b). Use of MEDLINE by physicians for clinical problem solving. *JAMA, 269*(24), 3124-9.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

McCray, A. T., Aronson, A. R., Browne, A. C., Rindflesch, T. C., Razi, A., & Srinivasan, S. (1993). UMLS knowledge for biomedical language processing. *Bull Med Libr Assoc, 81*(2), 184-94.

McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods Inf Med, 34*(1-2), 193-201.

MeSH. (1998). *Medical Subject Headings*. Bethesda (MD): National Library of Medicine.

Nelson, S. J., Olson, N. E., Fuller, L., Tuttle, M. S., Cole, W. G., & Sherertz, D. D. (1995). Identifying concepts in medical knowledge. *Medinfo, 8*(Pt 1), 33-6.

Nelson, S. J., Tuttle, M. S., Cole, W. G., Sherertz, D. D., Sperzel, W. D., Erlbaum, M. S., Fuller, L. L., & Olson, N. E. (1991). From meaning to term: semantic locality in the UMLS Metathesaurus. *Proceedings of Annual Symposium on Computer Applications in Medical Care (SCAMC)*, 209-13.

Platt, J. C. (1998). How to implement SVMs. *IEEE Intelligent Systems* (July/August), 26-28.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimitzation. In B. Scholkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in Kernal Methods* (pp. 185-208). Cambridge, Massachusetts: The MIT Press.

Salton, G. (1988). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Reading, Mass.: Addison-Wesley.

Salton, G. (1992). The State of Retrieval-System Evaluation. *Information Processing & Management, 28*(4), 441-449.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In E. A. Fox (Ed.), *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 138-146).

Schamber, L. (1994). Relevance and Information Behavior. *Annual Review of Information Science and Technology, 29,* 3-48.

Schuyler, P. L., McCray, A. T., & Schoolman, H. M. (1989). A test collection for experimentation in bibliographic retrieval. In B. Barber, D. Cao, D. Qin, & G. Wagner (Eds.), *MEDINFO 89* (pp. 810-912). Amsterdam: North-Holland.

Soergel, D. (1994). Indexing and Retrieval Performance: The Logical Evidence. *Journal of the American Society For Information Science, 45*(8), 589-599.

Sparck Jones, K. (1981). *Information retrieval experiment*. London; Boston: Butterworths.

Su, L. T. (1992). Evaluation Measures For Interactive Information-Retrieval. *Information Processing & Management, 28*(4), 503-516.

Tague-Sutcliffe, J. (1992). The Pragmatics of Information-Retrieval Experimentation, Revisited. *Information Processing & Management, 28*(4), 467-490.

UMLS. (1998). *UMLS Knowledge Sources* (9th ed.). Bethesda (MD): National Library of Medicine.

Wilbur, W. J. (1996). Human subjectivity and performance limits in document retrieval. *Information Processing & Management, 32*(5), 515-527.

Wilbur, W. J., & Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med, 26*(3), 209-22.

9. Appendix: Indexing Initiative Team Members

Alexa T. McCray, LHNCBC/OD

Nancy D. Wright, LO/Index

Harold M. Schoolman,[1] NLM/OD

Alan R. Aronson, LHNCBC/CgSB

Olivier Bodenreider, LHNCBC/CgSB

H. Florence Chang, LHNCBC/CgSB

Tamas E. Doszkocs, SIS

George (Mike) F. Hazard, SIS

William T. Hole, LHNCBC/CSB

Susanne M. Humphrey, LHNCBC/CSB

James R. Marcetich, LO/Index

James G. Mork, LHNCBC/CgSB

Stuart J. Nelson, LO/MeSH

Thomas C. Rindflesch, LHNCBC/CgSB

John M. Rozier, OCCS

Catherine R. Selden, LO/NICHSR

Sara J. Tybaert, LO/BSD

W. John Wilbur, NCBI

1. Dr. Schoolman has recently retired from NLM.