

## The NLM Indexing Initiative

Alan R. Aronson<sup>†</sup>, PhD, Olivier Bodenreider<sup>†</sup>, MD, PhD, H. Florence Chang<sup>†</sup>, MS,  
Susanne M. Humphrey<sup>†</sup>, MLS, James G. Mork<sup>†</sup>, MS, Stuart J. Nelson<sup>‡</sup>, MD,  
Thomas C. Rindflesch<sup>†</sup>, PhD, W. John Wilbur<sup>§</sup>, MD, PhD

<sup>†</sup>Lister Hill National Center for Biomedical Communications (LHNCBC),

<sup>‡</sup>Division of Library Operations (LO),

<sup>§</sup>National Center for Biotechnology Information (NCBI)

National Library of Medicine, Bethesda, MD 20894

*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. We describe the current state of the project and discuss our plans for the future.*

### INTRODUCTION

Human indexing is an expensive and labor-intensive activity. The total costs of indexing at the National Library of Medicine (NLM) include data entry, NLM staff indexing and revising, contract indexing, equipment, and telecommunications costs. Indexers are highly trained individuals, not only in MEDLINE<sup>®</sup> 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.

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. These considerations led to the instigation of the Indexing Initiative at the library. Automated methods developed and implemented within the project will have an important impact on NLM's ability to continue to provide high-quality services to its constituents.

### BACKGROUND

For more than 150 years, NLM 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<sup>®</sup> controlled vocabulary<sup>1</sup> and, since 1974, author abstracts of many 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.

Some goals and assumptions were made at the beginning of the project. First, our ultimate goal is better retrieval of biomedical information, not just better conformity to indexing rules and practices. Second, NLM's MeSH vocabulary and the UMLS<sup>®</sup> Knowledge Sources<sup>2,3</sup> will continue to exist and grow. And finally, free text in the form of titles and abstracts will continue to be available, but the full text of journal articles in electronic form will also become increasingly available.

Early IND efforts consisted of several projects devoted to the development of novel indexing methods and, in addition, projects for consideration of evaluation and policy issues.

## METHODS

### The Indexing Initiative System

Recent IND efforts have used the results of the initial projects and have focused on the creation of a system for exploring different ways of producing recommended indexing terms. The result is the IND System which 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.

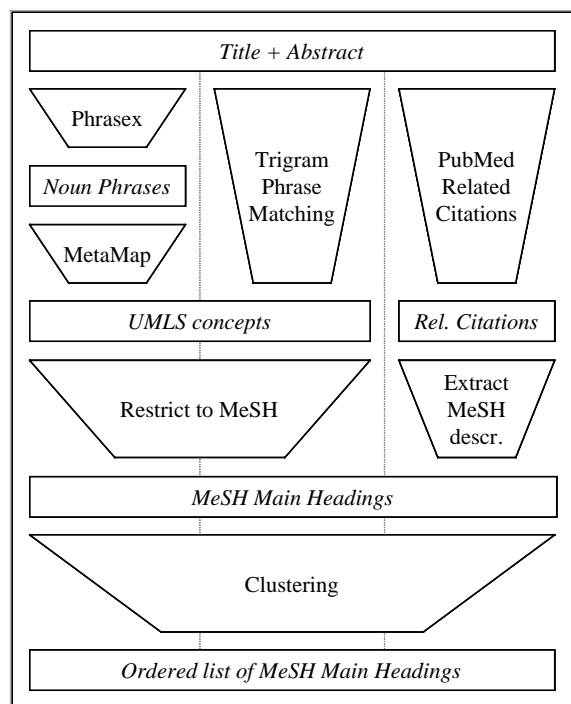


Figure 1. The Indexing Initiative System

The top portion of the diagram consists of three paths, or methods, for creating a list of recommended indexing terms: MetaMap Indexing, Trigram Phrase Matching, and PubMed Related Citations. The two left paths actually compute UMLS Metathesaurus® concepts which are passed to the Restrict to MeSH method. The results from each path are weighted and combined using the Clustering method. The system is highly parameterized not only by path weights but also by several parameters specific to the Restrict to MeSH and Clustering methods. We now describe the methods in the current IND System.

#### MetaMap Indexing

The MetaMap Indexing (MMI) method of discovering UMLS concepts consists of applying the MetaMap program<sup>4,5</sup> to a body of text and then ordering the resulting concepts using a ranking function. MetaMap finds Metathesaurus concepts in five steps:

1. Parsing: Arbitrary text is parsed into simple noun phrases using the SPECIALIST(tm) minimal commitment parser.<sup>4</sup>

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, inflectional variants and meaningful combinations of these.<sup>6</sup>
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 of the input phrase), variation, coverage and cohesiveness. The candidates are ordered according to mapping strength.
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.

Finally, MetaMap Indexing examines all the concepts assigned by MetaMap to a given citation and ranks them for how well they represent the content of the citation. The ranking function 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 which has the effect of giving title concepts overwhelmingly good rankings.

#### Trigram Phrase Matching

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.

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 internal punctuation.
2. For each phrase produced in 1, compute the similarity 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, the Trigram Phrase Matching algorithm produces UMLS concepts which are subsequently restricted to MeSH headings as described in the next section.

### Restrict to MeSH

The representation of meaning in the UMLS is organized according to the principle of semantic locality<sup>7,8</sup> 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.<sup>9</sup>

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.

The algorithm for restricting UMLS concepts to MeSH terms can be tuned from a strict mode (high precision) to a relaxed mode (high recall). The method that we use here is an intermediate mode between high precision and high recall, and appears to be optimal in the context of the Indexing Initiative, which ranks and clusters an array of indexing terms based on a range of methodologies.

### PubMed Related Citations

The PubMed Related Citations method directly computes a ranked list of MeSH headings based on a given title and abstract. 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. A list of 310 common, but uninformative, words (also known as stopwords) are eliminated from processing, and a limited amount of stemming of words is done; but no thesaurus is used in processing. When this method is used in PubMed, words are obtained from the title, abstract, and MeSH fields of MEDLINE citations. For indexing purposes, however, we use only the title and abstract.

Having obtained the set of terms that represent each document, the next step is to assign global and local

weights to each term. The global weight<sup>10</sup> is used in weighting the term throughout the database. The global weight of a term is greater for the less frequent terms. The local weight is  $\log(n+1)$  where  $n$  is the number of times the term occurs in a document. The product of the two weights is the weight of the term.

The similarity of two documents is computed using the term weights defined above and is an example of vector cosine scoring originated by Gerard Salton.<sup>11</sup> Our approach differs from other approaches in the way we calculate the local and global weights for the individual terms.

Recommended index terms are extracted from the MeSH fields of documents most similar to a given document.

### 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 play a role in that confidence: the method of finding the heading, how effectively the method found the heading, the location in the text of the nominal phrase that led to that suggestion, and the semantic consistency of the suggested heading with the other suggested headings.

The clustering algorithm embodies these principles in a formula which computes the rank score of each suggested indexing term. The formula uses term weights, estimates of the importance of the term based on where and how the term arose, and combines weights of related terms into a final rank score. The result of the clustering process is a ranked list of MeSH headings representing the combined recommendations of the constituent indexing methods.

### Parameter Tuning

The high degree of parameterization of the IND System allows us to test the components for their relative contribution to the results. We can, for example, compare the same method using different parameter settings or the same settings across different methods. We performed such experiments to determine optimal system parameter values 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 parameters. Recommended indexing was compared with the terms assigned by NLM indexers, and precision/recall values were computed. The experiments show that the MetaMap Indexing path is the single strongest path and that the Trigram Phrase Matching and PubMed Related Citations paths perform well as more recommended terms are considered. Further experiments combining paths showed that a combination of MetaMap Indexing and PubMed Related Citations gives results closest to the manually-assigned indexing.

## RESULTS

We now give sample results produced by the IND System. Consider the MEDLINE citation in Table 1,

showing the unique identifier (UI), title (TI), abstract (AB), and humanly assigned MeSH headings (MH).

UI	98018928
TI	Bupivacaine inhibition of L-type calcium current in ventricular cardiomyocytes of hamster.
AB	<p>BACKGROUND: The local anesthetic bupivacaine is cardiotoxic when accidentally injected into the circulation. Such cardiotoxicity might involve an inhibition of cardiac L-type <math>\text{Ca}^{2+}</math> current (ICa,L). This study was designed to define the mechanism of bupivacaine inhibition of ICa,L.</p> <p>...</p> <p>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.</p>
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

Table 1. A MEDLINE citation

The human indexing has nine terms, four of which come from a set of high frequency terms known as check tags (in this example: Animal; Hamsters; Male; Support, Non-U.S. Gov't). Table 2 shows, in rank order, some of the 125 recommended MeSH headings.

The System finds all five headings that are not check tags; these are shown in bold in Table 2. Note, however, that the rank score for "Dose-Response Relationship, Drug" is very low. With regard to check tags (excluded from the table), the System finds the check tags "Animal" and "Hamsters" but not the check tags "Male" and "Support, Non-U.S. Gov't".

Further analysis of the results shows that the System 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;
- "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

- "Heart Ventricle": The cardiomyocytes are taken from the heart ventricle;
- "Patch-Clamp Techniques": This method is also described in the abstract.

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

Table 2. IND System indexing (excluding check tags)

These results are typical: inclusion of most of the main headings plus additional relevant terms. Quantification of this observation, including the issue of cut-off points for acceptable indexing terms, remains for the experiments described in the next section.

## FUTURE WORK

Much of the current Indexing Initiative research focuses on improving the basic indexing methods. MetaMap Indexing is undergoing a major effort at introducing high-level tokenization for more accurate detection of acronyms and other special patterns in text. Several Machine Learning algorithms (naïve Bayes, adaptive boosting and support vector machines) are being tested to see if they can improve the PubMed Related Citations method. And recent research defining semantic proximity, a precise way of computing the "semantic distance" between a given pair of UMLS concepts, shows promise for quantifying the idea of semantic locality and thereby improving the Restrict to MeSH method.

In addition, we are about to embark on a major evaluation effort of the IND System focusing on retrieval performance rather than quality of indexing. The evaluation will include standard information retrieval

experiments using several test collections of MEDLINE citations. Other forms of evaluation will also be explored.

Finally, three research efforts extending the utility or scope of the research have begun or will commence in the near future.

### Word Sense Disambiguation

Error analysis performed during the evaluation process indicated word sense disambiguation as an area of focus for continued enhancement of the Indexing Initiative System. Indexing errors due to word sense ambiguity arise when the UMLS Metathesaurus has a single string referring to two or more distinct concepts. We do not currently have the means of choosing which concept is appropriate in the given textual context. Current research in statistically-based natural language processing addresses automatic resolution of this type of ambiguity.<sup>12</sup> 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<sup>13</sup> which minimizes this effort by first concentrating on non-ambiguous training text. In addition the work on Journal Descriptors described below offers another promising approach to word sense disambiguation.

### Full Text Processing

A second major area of planned research recognizes the fact that our current indexing methods rely only on titles and abstracts, while human indexers base their analysis on the full text of an article. This restriction causes the computer-generated terms to suffer recall errors in comparison to the humanly assigned document descriptors.

One approach to full text processing involves submitting all of the 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<sup>14</sup> is concerned with identifying key topics and sections in a full-text article. Additionally, insights from human indexer practice 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.

### Journal Descriptor (JD) Indexing

As a final area of research, we are investigating a novel approach to fully-automated indexing based on NLM's practice of maintaining a subject index to journal titles using terms, Journal Descriptors, corresponding to specialties associated with biomedicine.<sup>15,16</sup> JD Indexing draws its strength from its ability to associate JDs with a word, a phrase or, indeed, any body of text. Preliminary experiments indicate that this ability shows promise when applied to the word sense disambiguation problem.

### Acknowledgements

We would like to thank our project leader, Alexa T. McCray and the other Indexing Initiative team members, Tamas E. Doszkocs, George F. Hazard, William T. Hole, James R. Marcetich, Catherine R. Selden and Sara J. Tybaert for their essential contributions to this research effort.

### References

1. MeSH. *Medical Subject Headings*. Bethesda (MD): National Library of Medicine, 1999.
2. Lindberg DAB, Humphreys BL, and McCray AT. The Unified Medical Language System. *Methods of Information in Medicine*, 1993; 32(4): 281-91.
3. UMLS. UMLS Knowledge Sources (9th ed.). Bethesda (MD): National Library of Medicine, 1999.
4. Aronson AR, Rindfleisch TC, and Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94*, 1994: 197-216.
5. Aronson AR and Rindfleisch TC. Query expansion using the UMLS Metathesaurus. *Proceedings of AMIA Annual Fall Symposium*, 1997: 485-9.
6. Aronson AR. The effect of textual variation on concept based information retrieval. *Proceedings of AMIA Annual Fall Symposium*, 1996: 373-7.
7. Nelson SJ, Tuttle MS, Cole WG, Sherertz DD, Sperzel WD, Erlbaum MS, Fuller LL, and Olson NE. From meaning to term: semantic locality in the UMLS Metathesaurus. *Proceedings of Annual Symposium on Computer Applications in Medical Care (SCAMC)*, 1991: 209-13.
8. McCray AT, and Nelson SJ. The representation of meaning in the UMLS. *Methods of Information in Medicine*, 1995; 34(1-2): 193-201.
9. Bodenreider O, Nelson SJ, Hole WT, and Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proceedings of AMIA Annual Fall Symposium*, 1998: 815-9.
10. Wilbur WJ and Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Computers in Biology and Medicine*, 1996; 26(3): 209-22.
11. Salton G. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley, 1988.
12. Manning CD and Schütze H. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press, 1999.
13. Daelemans W. Memory-based lexical acquisition and processing. In Steffens P (Ed.), *Machine translation and the lexicon: 3rd International EAMT Workshop Proceedings*, 1993 Apr 26-28; Heidelberg, Germany. Berlin: Springer-Verlag, 85-98.
14. Lin C., and Hovy E. Identifying topics by position. *Proceedings of the Fifth Conference on Applied Natural Language Processing* (Association for Computational Linguistics), 1997: 283-290.
15. Humphrey SM. A new approach to automatic indexing using journal descriptors. *Proceedings of the ASIS Annual Meeting*, 1998; 35: 496-500.
16. Humphrey SM. Automatic indexing of documents from journal descriptors: A preliminary investigation. *Journal of the American Society for Information Science*, 1999; 50(8): 661-674.