# Automated Indexing of the Hazardous Substances Data Bank (HSDB)

**Carlo Nuss, MS, Hua Florence Chang, MS, Dorothy Moore, MS, George C. Fonger, BS**
**Specialized Information Services Division, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland**

Abstract: The Hazardous Substances Data Bank (HSDB), produced and maintained by the National Library of Medicine (NLM), contains over 4600 records on potentially hazardous chemicals. To enhance information retrieval from HSDB, NLM has undertaken the development of an automated HSDB indexing protocol as part of its Indexing Initiative. The NLM Indexing Initiative investigates methods whereby automated indexing may partially or completely substitute for human indexing. The poster's purpose is to describe the HSDB Automated Indexing Project.

The Hazardous Substances Data Bank (HSDB) is a factual data file produced by the National Library of Medicine (NLM) that contains over 4600 records on potentially hazardous chemicals. Each record is highly structured and includes up to 150 fields. In order to enhance the accuracy of information retrieval from HSDB, the Specialized Information Services (SIS) Division at NLM has undertaken the development, testing, and implementation of an automated indexing protocol for the data bank as part of the NLM Indexing Initiative.

The NLM Indexing Initiative is a project that investigates methods whereby automated indexing may partially or completely substitute for human indexing of biomedical information. The automated indexing system is based on the application of three fundamental methodologies: the MetaMap program, which maps text to concepts in the Unified Medical Language System (UMLS) Metathesaurus; the trigram phrase algorithm, which uses character trigrams to match text to Metathesaurus concepts; and a variant of the PubMed-related citations algorithm to find MeSH terms related to input text. The UMLS concepts from the first two methods are then mapped to MeSH main headings through the Restrict to MeSH algorithm. The resulting list of MeSH terms are then clustered into a ranked list of recommended indexing terms. The three methods can be used alone or in combination.

The HSDB Automated Indexing Project consists of four phases. Phase I – Human Review of Automated Indexing – includes selection of HSDB data fields most suitable for MeSH indexing; assessment of indexing accuracy; and determination of baseline indexing precision. Phase II – Algorithmic Tweaking to Increase Indexing Precision – includes identification of the indexing method or combination of methods yielding optimal results; determination of a cut-off point for ranked MeSH lists based on indexing precision; elimination of problematic or unnecessary terms; and addition of more relevant or necessary terms. Phase III – Test System Setup – includes automated indexing of all HSDB records; setup of a search system replica on a test server; and integration of MeSH indexing with HSDB data on the test server. Phase IV – Retrieval Evaluation – includes testing of retrieval by running strategic search queries; assessment of retrieval capability by comparing retrieval obtained from indexed versus nonindexed HSDB records; and usability testing with internal and external test groups.

The purpose of the poster is to present the challenges, process, and results of applying automated indexing methodologies to a data bank with highly structured records, a variety of text and data formats, and complex technical and biomedical terminology.