



Lister Hill National Center for

**Biomedical Communications**

An Intramural Research Division of the U.S. National Library of Medicine

## **A Report to the Board of Scientific Counselors**

### **April 2016**

### **NLM Medical Text Indexer**

Dina Demner-Fushman, MD, PhD  
Communications Engineering Branch

James Mork, MSc  
Cognitive Science branch

---

U.S. National Library of Medicine, LHNCBC  
8600 Rockville Pike, Building 38A  
Bethesda, MD 20894



U.S. National Library of Medicine



## Contents

Background.....	2
Objective.....	4
Significance.....	4
Methods and Procedures.....	4
Recent Enhancements.....	6
Vocabulary Density Filtering.....	7
Check Tag Selection.....	7
Subheading Attachment.....	8
International BioASQ Challenge.....	8
Usage and Application of MTI.....	8
MTI First Line (MTIFL).....	9
Semi-Automatic Indexing for a Select Set of 51 Journals.....	10
MeSH on Demand.....	10
OLDMEDLINE Mapping and Updating.....	10
Project Status.....	11
Ongoing Research Efforts.....	11
Focused Full Text.....	11
Learning to Rank MTI Suggestions.....	13
Evaluation Plan.....	13
Summary and Future Work.....	14
References.....	16

## List of Figures

Figure 1 MTI Precision vs Indexer Average Daily Use Between 2007 and 2015.....	3
Figure 2 MTI Processing Flow Diagram.....	5
Figure 3 Number of Unique MHs Used by the Number of Journals.....	6
Figure 4 Title and Abstract versus Full Text Example (PMID: 24000132).....	12
Figure 5 MTI Performance 2007 – Present (November 18, 2015).....	14
Figure 6 Current Indexing vs Suggested Relation Extraction Example.....	15

## List of Tables

Table 1 Machine Learning MeSH Terms and Performance Improvements.....	8
Table 2 MTI Tool Usage for 2010 - 2015 by Category.....	9
Table 3 Results of the Evaluation of Check Tags Assignment Based on Full Text Articles.....	13

## Background

Ever since John Shaw Billings published “*Specimen Fasciculus of a Catalogue of the National Medical Library*” in 1876<sup>1</sup>, the US National Library of Medicine (NLM) has provided access to the biomedical literature through the analytical efforts of human indexers. Since 1960, indexers have manually assigned terms from the Medical Subject Headings (MeSH<sup>®</sup>) controlled vocabulary to describe and identify key points in an article. Search engines have advanced to where they are now capable of automatically indexing the full text of scientific articles, however, significant improvements in retrieval effectiveness are still obtained by combining manual and automated indexing [1].

127 indexers and revisers at the NLM provide manual indexing to over 800,000 citations a year, covering 5,618 international biomedical journals in 37 languages using MeSH<sup>2</sup>. 2016 MeSH consists of 27,883 Descriptors (Main Headings) and over 230,000 Supplementary Concept Records<sup>3</sup>. Each MeSH Descriptor can also be further qualified by selections from a collection of 82 topical Subheadings.

In the face of a growing workload and dwindling resources, the NLM launched the Indexing Initiative project [2] in the mid-1990s. This cross-library team’s mission was to explore indexing methodologies to help ensure that MEDLINE<sup>®</sup> and other NLM document collections maintain their quality and currency and thereby contribute to the NLM mission of maintaining quality access to the biomedical literature. The NLM Medical Text Indexer (MTI) [3], a tool for analyzing and characterizing documents automatically, is the main product of this project. In this report, we will refer to MTI functions as automated indexing, using the sense of the term that pertains to deriving documents’ characteristics and main points automatically, similar to automated abstractive summarization. One of the uses for this automatically derived document indexing by MTI is to provide assisted indexing to the NLM indexers.

The NLM Medical Text Indexer has been providing indexing assistance since 2002 by combining the expertise of indexers working at NLM with natural language processing technology to annotate the biomedical literature with Medical Subject Headings more efficiently and consistently. MTI continues to improve its performance, expand its use, and advance research in automated indexing since its inception. As with many other NLM projects, an initiative that started as a practical application to address immediate needs has developed into a rich research project that provides a platform for and successfully advances informatics research, while continuing to be a useful and widely used tool.

Figure 1 shows how MTI has nearly doubled the precision of recommendations made to the indexers since 2007; correspondingly, adoption by the indexers and confidence in MTI have also increased, leading to an expansion of the role MTI has in NLM Indexing:

- MTI First Line (MTIFL) indexing, in which MTI is treated as a novice indexer, has expanded to 386 high performing journals making up approximately 10% of the yearly indexing,
- MTI has become the primary indexer for the semi-automatic indexing of a select set of 51 journals that perform as well as the MTIFL journals, but, require more specialized filtering,
- MTI was used to provide automatic mapping and updating of the indexing for over two million OLDMEDLINE citations. OLDMEDLINE are citations from 1946 to 1965 that were originally printed in hardcopy indexes published prior to 1966. MTI will continue to be used for mapping and updating the indexing for OLDMEDLINE citations as they are added to PubMed.

---

<sup>1</sup> <https://apps.nlm.nih.gov/175/milestones.cfm>

<sup>2</sup> [https://www.nlm.nih.gov/bsd/bsd\\_key.html](https://www.nlm.nih.gov/bsd/bsd_key.html)

<sup>3</sup> <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>

MTI is in daily use to assist Indexers, Catalogers<sup>4</sup>, and NLM's History of Medicine Division (HMD)<sup>5</sup> in their indexing efforts. Every weeknight MTI provides recommendations for approximately 4,000 new citations for Indexing and processes a mixed file of approximately 7,000 old and new records for both Cataloging and HMD. MTI was also used on a regular basis between 2002 and 2012 to provide fully-automated keyword indexing for the NLM Gateway Meeting Abstract collection<sup>6</sup>, which was not manually indexed. MTI is periodically used during the indexing year to re-process the nearly 350,000 In-Process citations whenever sufficient improvements have been made to the MTI algorithms, making sure indexers receive the best possible recommendations.

We have also extended the reach of MTI beyond NLM Indexing by developing in collaboration with the MeSH Section at NLM, an easy to use keyword identification interface to MTI called *MeSH on Demand*<sup>7</sup>. MTI also continues to be a rich platform for research providing the impetus and baselines for the International BioASQ challenges<sup>8</sup> on biomedical semantic indexing over the last three years, providing avenues of research into ways machine learning may help to improve automated indexing, and research into the strategic use of full text articles as they become available.

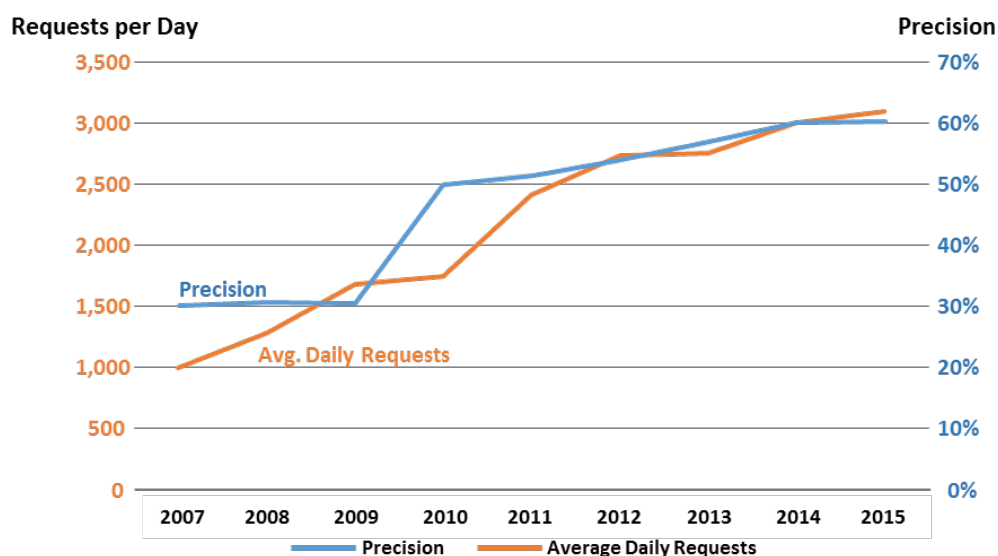


Figure 1 MTI Precision vs Indexer Average Daily Use Between 2007 and 2015

MTI was previously presented in the 1999, 2008, and 2012 *Indexing Initiative* reports to the Board of Scientific Counselors and in the 2002 *Automated and Semi-automated Indexing* report to the Board of Regents. In this report, we briefly present the overall architecture of the MTI system and its performance in comparison to the human indexers. We primarily focus on MTI research in biomedical semantic indexing (interchangeably used as a synonym for automated indexing in this report).

<sup>4</sup> <http://www.nlm.nih.gov/tsd/cataloging/mainpge.html>

<sup>5</sup> <http://www.nlm.nih.gov/hmd/index.html>

<sup>6</sup> <http://www.nlm.nih.gov/pubs/factsheets/gateway.html>

<sup>7</sup> <https://www.nlm.nih.gov/mesh/MeSHonDemand.html>

<sup>8</sup> <http://www.bioasq.org/>

## Objective

The objective of the NLM Medical Text Indexer project is to investigate methods for automatic indexing to enhance access to NLM document collections including MEDLINE. MTI success is measured in contributions to semantic indexing research and increases in indexing efficiency and consistency while maintaining and improving access to biomedical information.

## Significance

Text understanding beyond simple keyword matches is an increasingly important language processing task, and automated indexing dedicated to finding main points of a given text is an integral part of this research. Manual indexing is a labor-intensive and expensive activity with indexers completing more than 806,000 citations in FY 2015<sup>9</sup> at an average cost of approximately \$9.40 per article. Considerations such as the increasing demand on the NLM indexing resources and staff coupled with the flat budgets seen throughout federal agencies make clear that if (semi-) automated methods can be successfully developed and implemented, the project will have a considerable impact on the ability of NLM to continue to provide high-quality services to its patrons. Equally importantly, the project should continue to contribute to information science research promising future performance improvements, as well as offer training opportunities to young researchers in the field.

Over the years, the Indexing Initiative project, headed by Dr. Alan Aronson through 2014, provided an excellent training venue for young researchers as well as numerous in-house and worldwide collaborative opportunities. Fine examples of such training/collaborative work are subheading attachment and the selection of supervised machine learning algorithms – two major MTI enhancements that are the direct result of research with post-doctoral fellows. Research on the vocabulary density filtering was inspired in part by one of our indexers. This is but one illustration of the benefits of the multidisciplinary approach taken by NLM from the inception of the project: the subject matter experts are interested in advancing automated indexing research that supports their daily tasks. Over the last three years MTI has also been the cornerstone of the international BioASQ Challenge which is stimulating international and corporate research on systems directly related to our work.

## Methods and Procedures

Given any biomedical text, MTI produces a ranked list of controlled vocabulary terms (MeSH) that summarizes the main points of the text using MeSH Main Headings (MH), Subheadings (SH), Check Tags (CT), and Supplementary Concept Records (SCRs).

MHs are the main descriptors or headings from the MeSH controlled vocabulary (e.g., *Lung*). SHs are used to clarify or qualify aspects of the MHs (e.g., *Lung/abnormalities* means that the article is about *abnormalities* associated with the *Lung* more than the *Lung* itself), and CTs are a special type of MHs that are required to be included for each article and cover species, sex, human age groups, historical periods, pregnancy, and various types of research support (e.g., *Male*). SCRs are used to index chemicals, drugs (e.g., *thiozamin*), and other concepts such as diseases for which no descriptor exists (e.g., *Myeloperoxidase Deficiency*)<sup>10</sup>.

---

<sup>9</sup> [https://www.nlm.nih.gov/bsd/bsd\\_key.html](https://www.nlm.nih.gov/bsd/bsd_key.html)

<sup>10</sup> [https://www.nlm.nih.gov/mesh/intro\\_record\\_types.html](https://www.nlm.nih.gov/mesh/intro_record_types.html)

Multiple methods were investigated during the initial phases of the Indexing Initiative project and eventually a voting-style mechanism incorporating two of the best performing experimental systems (MetaMap Indexing and PubMed Related Citations) was developed into the NLM Medical Text Indexer (MTI). MTI was later expanded to take advantage of supervised machine learning for a small number of frequently used indexing Check Tags.

Routine MTI processing involves receiving a daily XML formatted MEDLINE file<sup>11</sup> which contains a list of Completed, In-Process, and In-Data-Review citations. All processing is done offline, and the MTI results are then stored in a database for later use by the indexers. Pre-loading of the MTI results provides the indexers immediate access to the results without delaying their work. Figure 2 depicts the processing flow as MEDLINE citations are processed through the various components of the MTI system. Each of the major MTI components is described briefly below.

**MetaMap Indexing (MMI)** [4]: a method that applies a ranking function to UMLS Metathesaurus concepts identified by MetaMap emphasizing concepts found in the title of the citation [5].

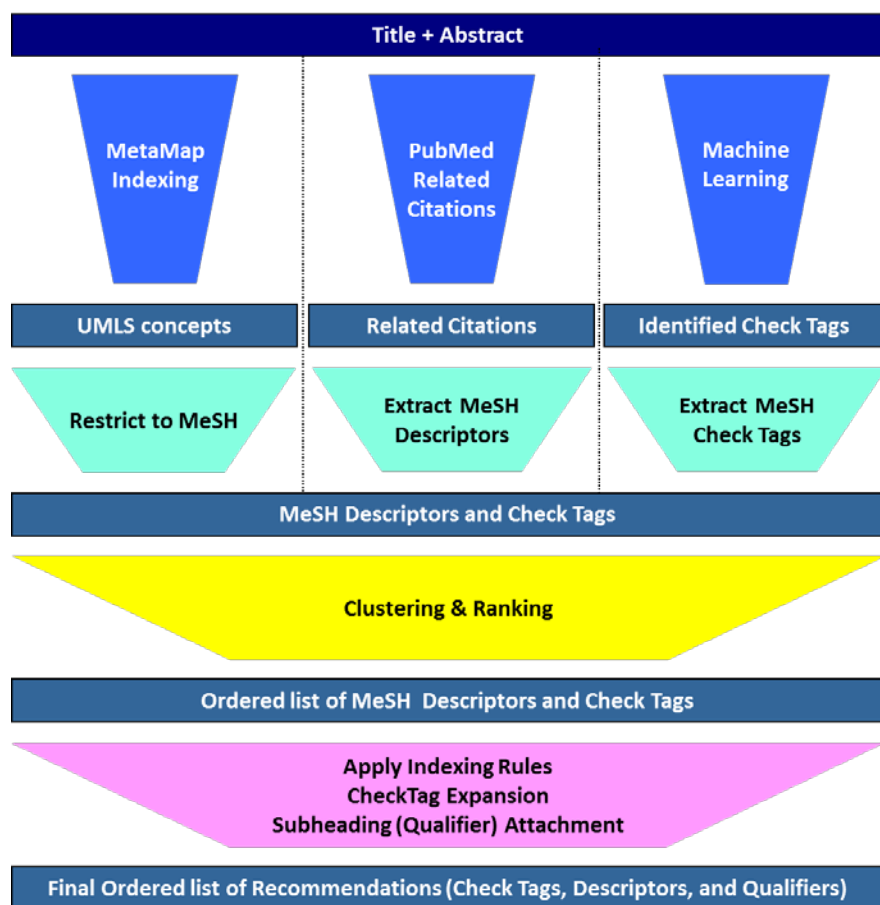


Figure 2 MTI Processing Flow Diagram

**PubMed Related Citations (PRC)** [6]: MTI currently has two methods available for determining PubMed Related Citations for the text it is processing. If MTI is working with a MEDLINE citation we use

<sup>11</sup> [http://www.nlm.nih.gov/bsd/licensee/elements\\_descriptions.html](http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html)

the list of PRC pre-defined by the PubMed system<sup>12</sup>. If MTI is processing free form text, MTI will default to using the in-house TexTool<sup>13</sup> implementation of PRC.

**Machine Learning [7,8]:** we train our AdaBoostM1 classifier on the last three years of human indexing in an effort to improve performance on a small set of the most frequently used terms in MeSH that have showed improvements during testing. More detailed information on machine learning in MTI is provided in the *Check Tag Selection* section later in this report.

**Restrict to MeSH [9]:** a method which maps UMLS® Metathesaurus® concepts to their closest set of MeSH Headings using the Restrict to MeSH mapping file.

**Extract MeSH Descriptors:** retrieving the MeSH Headings from the PRC in MEDLINE format and tracking whether the MeSH Heading is a major descriptor (starred) or not.

**Clustering and Ranking [10]:** the ranked lists of MHs produced by the methods described so far must be aggregated 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.

**Post-Processing:** once all of the recommendations are ranked and selected, validation of the recommendations is done based on the end-user requirements. Typically, CTs are added based on triggers from the text, recommendations are added or removed based on the specific journal using our Vocabulary Density Filtering, and then finally MTI performs subheading attachment to individual headings and for the text in general. Not all citations processed by MTI go through all of the components listed above. MTI has various filtering levels and special handling rules that enable different processing pathways.

## Recent Enhancements

In this section we detail a few of the recent enhancements that have made a major impact on MTI performance and expanded its usefulness to the indexers.

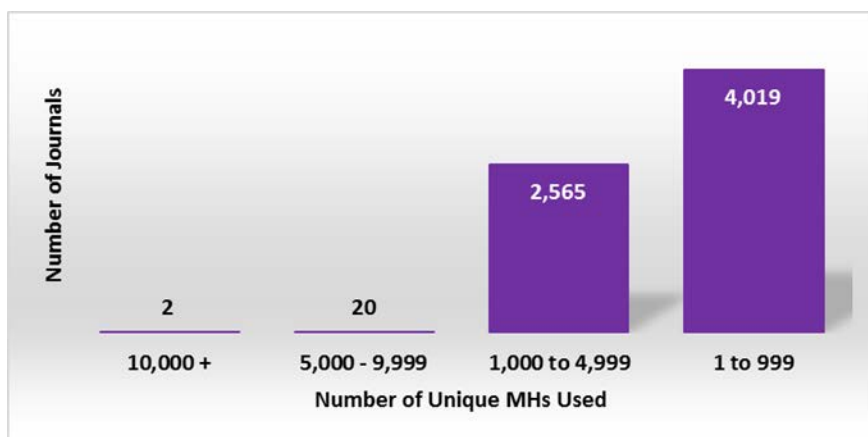


Figure 3 Number of Unique MHs Used by the Number of Journals

<sup>12</sup> <http://www.nlm.nih.gov/pubs/factsheets/pubmed.html>

<sup>13</sup> <http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/Textool/>

### Vocabulary Density Filtering

Our earlier study [11,12] found that 83.81% of the used MHs are assigned to only 500 or fewer journals and 271 MHs are assigned to only a single journal. The study also discovered that, on average, only 3.68% (999) of the 27,149 terms in the 2014 MeSH Vocabulary were used to describe the articles in journals. The actual distribution of unique MHs assigned to the 6,606 journals is illustrated in Figure 3. The Vocabulary Density filtering uses this MeSH Term frequency information from the last five years of indexing for each individual journal to help determine MTI recommendations that should either be removed or added prior to presenting the results to the indexers. The idea is that if articles in a journal have not been assigned a specific MeSH Term in the last five years, and it is not a new MeSH Term, it is not likely that it would be a good term for MTI to recommend for this new article. On the other hand, if 90% of the articles in a journal are always assigned a specific MeSH Term, it is reasonable to assume that this new article is likely to need that term as well.

The selective use of MHs implementing this simple approach leads to a 2.69 (+4.44%) improvement in Precision, 0.05 (+0.08%) increase in Recall, and a 1.36 (+2.23%) increase in the overall F<sub>1</sub> score.

### Check Tag Selection

Because MeSH indexing can be viewed as a multi-label categorization task, we use machine learning in an effort to improve both Recall and Precision on twelve Check Tags which are some of the most frequently used terms in MeSH [7,8]. There are some problems to consider when applying machine learning to MeSH indexing [13,14]. While not all of these are issues specific to machine learning, they will all have some effect on the training and ultimately the results. In an effort to mitigate some of these issues, we retrain the algorithm yearly, using the last three years of indexing to ensure close adherence to current indexing policy and to help reduce the chance of drift due to changes in MeSH over time. Factors include:

- There is an imbalance between positive and negative instances because MEDLINE data contains many more negative examples than positive ones for each MeSH heading,
- Even if a MeSH heading is correctly identified for a citation it might not be significant enough to be included in the indexing,
- There are natural variations in indexing [15], and
- Changes are made to indexing policy over time.

After experimenting with several approaches including Naïve Bayes, Rocchio, and AdaBoostM1, we found significant improvement for the twelve Check Tags listed in Table 1 using AdaBoostM1 with C4.5 as the base learner. AdaBoostM1 is an ensemble learning algorithm which samples iteratively from the training data according to the performance of a base learner. Table 1 shows the Check Tag, MTI F<sub>1</sub> scores prior to and after implementing the machine learning algorithms, and how much of an improvement is obtained for each Check Tag. In 2015, these twelve Check Tags made up almost 20% of the total number of MeSH terms assigned for the year – suggesting that any improvements to these Check Tags will significantly help MTI performance.

Check Tag	F <sub>1</sub> prior to ML	F <sub>1</sub> with ML	Improvement
Adolescent	0.2475	0.4236	+0.1761
Adult	0.1949	0.5684	+0.3735
Aged	0.1172	0.5467	+0.4295
Aged, 80 and over	0.0150	0.3089	+0.2939
Child, Preschool	0.0611	0.4540	+0.3929
Female	0.4606	0.7384	+0.2778
Humans	0.7998	0.9133	+0.1135
Infant	0.3439	0.4469	+0.1030



Male	0.3847	0.7114	+0.3267
Middle Aged	0.0101	0.5950	+0.5849
Swine	0.7104	0.7475	+0.0371
Young Adult	0.0283	0.3163	+0.2880

Table 1 Machine Learning MeSH Terms and Performance Improvements

### Subheading Attachment

The final MTI step in creating indexing recommendations is to perform subheading attachment. Subheading attachment is currently only done for the Indexers because Cataloging and HMD do not utilize subheadings from MTI. Subheadings are not attached to every MH recommended by MTI; the subheading attachment algorithms use several linguistic and statistical methods to determine what is appropriate for each MH based on the text and which subheadings are allowable for each MH [16-18]. A sample post-processing rule is: "If the main heading *Mutation* and a <DISEASE> term (from MeSH "C" Diseases Tree) appear in the indexing recommendations, then the pair <DISEASE>/*genetics* will also be recommended." Similar to the Vocabulary Density Filtering, we use historical data for MH/SH combinations to remove unlikely SHs and add SHs that are traditionally found with a given MH. MeSH specifies a subset of the subheadings that are allowed for each MH, so the subheading attachment algorithms utilize these rules to ensure that non-allowed combinations are not recommended by MTI. Based on the results of two user-centered studies, at most three subheadings are attached to each MH.

### International BioASQ Challenge

MTI has provided baselines every year for the international BioASQ Challenge since its inception in 2012. BioASQ is an ongoing series of challenges in biomedical semantic indexing and question answering with the aim of advancing the state of the art in efficient retrieval of biomedical text. The MTI indexing results are providing two of the baselines used in the "large-scale online biomedical semantic indexing" part of the challenge, which is designed to advance automated indexing of the most salient points of biomedical publications. The BioASQ Challenge evaluation of approaches to biomedical semantic indexing provided a continuous assessment of the indexing suggestions that are automatically generated by the MTI system used in support of the MEDLINE indexing process at the NLM. The benefits of participating in this community-wide evaluation for MTI have been two-fold: firstly, MTI was rigorously compared to systems developed by a world-wide community of researchers and industrial teams (e.g., *Aristotle University of Thessaloniki, Greece; University of Massachusetts Amherst, USA; Fudan University, China; and Transinsight GmbH, Germany*) all performing the same task; and secondly, the free exchange of the methods and ideas allowed the MTI team to incorporate the best practices explored by the participating teams. Incorporating some of these approaches into the MTI workflow in 2013-2014 improved the precision of MTI indexing suggestions by 4.44%.

### Usage and Application of MTI

MTI is used by in-house staff as well as researchers and institutions (e.g., *American Academy of Family Physicians (AAFP)*) from around the world every day.

We have worked over the years to simplify and expand access to MTI and the increased usage over the years shows that users are taking advantage of the various access methods. We briefly detail the various ways users can access MTI and Table 2 shows how much users have accessed MTI over the last five years.

- **Batch:** Web page that allows users to submit larger sets of documents to be processed via our distributed Computing Environment. The text is uploaded to our distributed computing environment for processing, and the user is notified by email when the batch has completed processing so they can download the results. [http://ii.nlm.nih.gov/Batch/UTS\\_Required/mti.shtml](http://ii.nlm.nih.gov/Batch/UTS_Required/mti.shtml)

- **API:** Downloadable Java-based Application Program Interface (API) Library that allows users to programmatically access all of our Indexing Initiative tools including MTI. The API library can be run on its own or incorporated into the user's own Java programs. Typically, the API is used to programmatically submit text to our Batch facility and receive the results back from within their own program. [http://ii.nlm.nih.gov/Web\\_API/index.shtml](http://ii.nlm.nih.gov/Web_API/index.shtml)
- **Interactive:** Web page that allows users to provide a small amount of text and to experiment with the full range of MTI options before submitting large quantities of text to be processed by MTI. <https://ii.nlm.nih.gov/Interactive/mti.shtml>
- **MeSH on Demand:** A web-based simplified interface to the MTI system that allows users to provide any text as input for MTI to summarize into a list of relevant MeSH terms. We discuss MeSH on Demand in more detail later in the report. <https://www.nlm.nih.gov/mesh/MeSHonDemand.html>

Table 2 shows MTI use numbers for 2010 through 2015 by each of the access categories for both the number of requests and the number of documents processed. For the Batch and API methods, we mostly see multiple documents, typically a set of MEDLINE citations to be processed for each request. For Interactive and MeSH on Demand, each request is considered a single document to be processed.

	Batch		API		Interactive	MeSH on Demand
	Requests	Documents	Requests	Documents		
<b>2010</b>	488	488	10	10	134	
<b>2011</b>	1,301	1,301	42,567	42,567	3,940	
<b>2012</b>	910	15,929,727	40,033	140,055	4,729	
<b>2013</b>	1,366	44,501,379	38,248	962,728	4,138	
<b>2014</b>	1,798	64,098,419	73,956	1,905,466	12,781	225,750
<b>2015</b>	1,585	49,289,778	42,623	13,705,423	37,258	216,309
	7,448	173,821,092	237,437	16,756,249	62,980	442,059

Table 2 MTI Tool Usage for 2010 - 2015 by Category

MTI also provides us with a tunable framework, in which we are able to customize the filtering and output to satisfy various use cases without having to create different versions of the program. This flexibility allows us to build customized uses of MTI rapidly for the operations side of NLM such as MTI First Line Indexing, Semi-Automatic Indexing for a Select Set of 51 Journals, MeSH on Demand, and OLDMEDLINE Mapping and Updating which are all detailed in the following subsections.

### MTI First Line (MTIFL)

The standard indexing process consists of two steps: 1) indexers assign MeSH to describe the content of an article based on a review of the full text, and 2) in-house revisers, senior staff who are expert indexers, review and modify the indexing and release it for searching and viewing in PubMed. MTIFL partially automates the indexing process by using MTI for the first step of indexing, focusing on only the titles and abstracts. In-house revisers continue with the second step, reviewing the MTIFL indexing using the full-text to add or delete MHs, and releasing the final indexing to PubMed. This human curation of MTIFL results is called MTIFL Completion. In February 2011, fourteen journals were initially selected to be included in the MTIFL pilot and 400 have been added since then for a total of 414 journals making up almost 10% of annual indexing. Journals selected for the MTIFL program must show above average MTI performance and low potential for manually created chemical flags (request to add a newly identified

chemical to MeSH) and GeneRIFs (Gene Reference into Function)<sup>14</sup> that are usually added by the indexer. In the case of MTIFL, the burden of creating the chemical flags and GeneRIFs would then shift to the reviser, which would be time consuming and undesirable.

### Semi-Automatic Indexing for a Select Set of 51 Journals

In 2015, MTI became the primary indexer for a select set of 51 journals where the MTI performance was found to be significantly higher than the average. This set of journals requires additional specialized MTI filtering focused more towards Precision, above the threshold for the MTIFL journals. For fifteen of the 51 journals, MTI also provides preliminary Index Medicus (IM), so-called, starred terms that are the main point of the article, further assisting the indexer. In 2014, prior to selection, these 51 journals collectively had an MTI average  $F_1$  of 0.6492, which was much higher than the overall  $F_1$  of 0.5807. Currently these journals are performing even better with an  $F_1$  of 0.8642, well above the average current MTI processing  $F_1$  of 0.5878. The MTI indexing for these journals is reviewed by a human reviewer to remove any erroneous MTI terms, add any important terms that MTI may have missed, and add Publication Types. These journals accounted for 15,509 articles in 2015 and will average approximately 12,000 articles a year going forward.

### MeSH on Demand

MeSH on Demand is a simplified interface to the MTI system. The MeSH on Demand interface maintained by NLM MeSH Section allows users to provide any text (e.g., MEDLINE citation or free text) as input and provides a list of relevant MeSH Descriptors and MeSH Supplementary Concepts that summarizes the input text and a list of the top ten citations related to the text in PubMed as a result. These results are heavily filtered in favor of terms with high confidence. The MeSH on Demand Tool is freely available to users at <http://www.nlm.nih.gov/mesh/MeSHonDemand.html>. MeSH on Demand was released to the public in February 2014 and averages approximately 18,000 hits per month. The MeSH on Demand Tool is an excellent starting point in searching PubMed for users who may not know the scientific nomenclature as recommended by institutions like *David D. Palmer Health Sciences Library*, *Tufts University*, and the *Alzheimer's Association Green-Field Library*. Two universities are evaluating the MeSH on Demand Tool to identify keywords for their course descriptions and lecture notes for easier retrieval by students and faculty. Several institutions, journals, and publishers, for example, *University at Buffalo Libraries*, have also provided authors with links to the MeSH on Demand Tool for use when they are adding keywords to their papers. MeSH on Demand was developed in close collaboration with the MeSH Section at NLM.

### OLDMEDLINE Mapping and Updating

There are 2,010,941 OLDMEDLINE<sup>15</sup> article citations from international biomedical journals that cover the fields of medicine, preclinical sciences and allied health sciences from 1946 to 1965 that were originally printed in hardcopy indexes published prior to 1966. The mapping project mapped the original subject headings assigned to the OLDMEDLINE citations when they appeared in the print indexes to the current MeSH vocabulary<sup>16</sup>. In 2015, MTI was used to update the indexing for each of the OLDMEDLINE citations using terms that were added to MeSH after these citations were originally indexed. Over 5 million new MTI-generated MeSH Terms were automatically added to the OLDMEDLINE articles, more than doubling the number of indexing terms for this set of citations. A good example of MTI updating the OLDMEDLINE indexing is PMID: 14312962, for which only the title is available. Given the title “*BACTERIO-CLINICAL COMPARISONS. VALUE OF THE PRACTICAL STUDY OF TRUE*

---

<sup>14</sup> <http://www.ncbi.nlm.nih.gov/gene/about-generif>

<sup>15</sup> PubMed Query: jsubsetom

<sup>16</sup> [https://www.nlm.nih.gov/databases/databases\\_oldmedline.html](https://www.nlm.nih.gov/databases/databases_oldmedline.html)

*POLYRESISTANCE IN PULMONARY TUBERCULOSIS PATIENTS WHO ARE CARRIERS OF MULTIPLE MONORESISTANCE. DEDUCTIONS AND THERAPEUTIC COMPARISONS*”, MTI added the new MeSH Term *Tuberculosis, Pulmonary*, which was not available in MeSH until 1999, to the pre-existing indexing: ANTITUBERCULAR AGENTS; DRUG RESISTANCE, MICROBIAL; and DRUG THERAPY.

## Project Status

MTI has provided indexing recommendations to the indexers for over thirteen years. It is, therefore, a mature project, but it is also very much a growing research effort. We are continually looking at ways to improve and expand the use of automated indexing. The success of MTI is grounded in the continuing collaboration between the indexers and the MTI project, so much so that we have a mechanism in place allowing indexers to provide immediate feedback for each article letting us know where MTI missed or incorrectly recommended terms. Through this close collaboration and improvements to MTI, the indexers have come to trust and better understand how MTI can assist them in their work. This has encouraged and expanded the use of MTI in other NLM Library Operations projects such as the MTIFL and OLDMEDLINE updating projects, as well as direct support for the public through programmatic and graphic user interfaces.

We plan to encourage research interest in this area through continuing support for the International BioASQ Challenge and LHC training program. MTI has gained international exposure as well as a rich forum for collaboration and research directly related to MTI through BioASQ. Over the years, numerous summer interns, Library Associates, postdoctoral fellows and visiting scientists were trained and contributed to MTI research directly or through related Indexing Initiative research.

## Ongoing Research Efforts

### Focused Full Text

For contractual reasons full text of the articles is not currently available to MTI, but, is used by the indexers to do their work. As full text becomes more available we are preparing to exploit the additional information provided in the full text of an article. Earlier work with full text [19,20] has shown that MTI is still able to identify key points of the article without adding irrelevant recommendations even with the added complexity and abundance of information available in the full text of an article. The earlier work however failed to show conclusively that processing the full text would improve MTI performance.

Over the years, indexers have noted that specific detailed information about the study subjects that MTI missed was in fact included in the full text of an article, but, not in the title or abstract. So, we have started exploring full text using a more focused approach looking for specific triggers rather than having MTI process the entire full text of an article. We preprocess the full text looking for specific triggers. We are currently researching simple methods for extracting specific Mice and Rat strains, species, age groups, and gender from the full text of publications.

<b>Title and Abstract (PMID: 24000132):</b>
<b>Effect of high sustained +Gz stress on myocardial mitochondrial ultrastructure, respiratory function, and antioxidant capacity in rats.</b>
Exposure to high sustained positive acceleration (+Gz) is known to have a pathophysiological effect on the heart of the rat. As critical regulators of cardiac myocyte survival and death, mitochondria may be crucially involved in +Gz-induced pathogenesis. It was, therefore, of interest to investigate myocardial mitochondrial ultrastructure, respiratory function, and antioxidant capacity in rats after exposure to +10 Gz for 5 min. The results showed that high +Gz stress could damage mitochondrial ultrastructure; this was apparent from swollen, degenerated, and reduced mitochondria, and mitochondrial cristae broken or disappeared. This resulted in significant changes of quantitative indicators of mitochondria morphometry, for example increased surface density, volume density, average volume, and average surface area, and reduced numerical density. The studies also revealed that exposure to +Gz stress induced dysfunction of the mitochondrial respiratory chain, reduced the activity of antioxidant enzymes (catalase, superoxide dismutase, and glutathione peroxidase), and increased malondialdehyde content. We thus conclude that high +Gz stress not only damaged mitochondrial ultrastructure but also impaired respiratory function and antioxidant capacity.
<b>Full Text "Materials and methods" section:</b>
Thirty male Sprague-Dawley rats, weighing 280 ± 20 g ...

Figure 4 Title and Abstract versus Full Text Example (PMID: 24000132)

This specific information tends to be found in the Methods section of the full text where the authors describe how their experiments were structured. Usually this is where we see information on the type of experiment subjects (*Animal, Humans*, or both), sex of the subjects (*Male* or *Female*), age of the human subjects (*Infant, Newborn; Infant; Child, Preschool; Child; Adolescent; Young Adult; Adult; Middle Aged; Aged; and Aged, 80 and over*), and if an Animal study, what kind of animals (*Mice, Rats, Hamsters*, etc.). A simple example of this can be seen in Figure 4 where we have highlighted the descriptions of the experiment subjects in the Title, Abstract, and Full Text. For PMID 24000132, Figure 4 illustrates how the author provided only the general description of "rats" for the experiment subjects in the Title and Abstract and nothing about what sex the rats were, or what specific type of rats they were. The full text on the other hand includes specific information in the "Materials and methods" section of the paper letting us know the subjects were "Male" "Sprague-Dawley rats" in the experiment. This information from the full text is critical to MTI because recommending just *Rats* would only provide one-third of the correct answer. The human indexer would use *Male; Rats; and Rats, Sprague-Dawley*.

Our study [21] showed that the full text of biomedical articles has potential to significantly improve automatic indexing of MEDLINE citations with MeSH headings pertaining to the study subjects' characteristics. Furthermore, we showed that simple rule-based methods significantly outperform the current automated indexing provided by the NLM Medical Text Indexer for the following headings: *Adolescent; Aged; Aged, 80 and over; Middle Aged; Young Adult; Female; Male; and Chick Embryo*. These encouraging results indicate we should continue exploring how to better use the full text for automated indexing of MEDLINE citations.

Table 3 presents the results of the evaluation of full text either added to the titles and abstracts or directly contributing Check Tags from our list of subject terms. The type, Recall, Precision, and F<sub>1</sub>, are provided for each experiment.

Experimental set-up	Recall	Precision	F <sub>1</sub>
MTI baseline (currently in use at NLM)	74.09%	81.35%	77.55%
Title expansion with sentences from the Methods section	78.19%	77.57%	77.88%
Abstract expansion with sentences from the Methods section	78.19%	78.26%	<b>78.20%</b>
Title expansion with sentences from the Methods section and captions	79.70%	76.05%	77.84%
Abstract expansion with sentences from the Methods section and captions	79.60%	76.79%	78.17%
Title expansion with sentences anywhere in the paper body	85.70%	58.07%	69.23%
Abstract expansion with sentences anywhere in the paper body	85.52%	59.86%	70.42%
Direct assignment of Check Tags with sentences from methods and captions	79.28%	74.17%	76.64%
Direct assignment of Check Tags with sentences anywhere in the paper body	86.42%	55.97%	67.94%

**Table 3 Results of the Evaluation of Check Tags Assignment Based on Full Text Articles**

### **Learning to Rank MTI Suggestions**

The best performing teams in the BioASQ evaluations showed that learning-to-rank algorithms [22,23] trained on the last five years of indexed PubMed citations and the candidate lists of MeSH terms generated using binary SVM classifiers, MTI recommendations and related citations have potential to improve recall and precision significantly at the cutoff levels corresponding to the numbers of recommendations (around 15) for a given citation [24,25]. We were able to replicate these results in a production-strength system building on work presented in [23,25] and are exploring additional features as well as learning the cutoff levels for suggestions.

Even at this early stage, Learning to Rank (L2R) promises to improve at least two currently underperforming areas. Using our MTI Test Collection, we found that L2R is helping with these issues:

1. Distinguishing whether to recommend a Publication Type or its indexing equivalent MeSH heading (e.g., *Clinical Study* Publication Type versus *Clinical Studies as Topic* MeSH heading). L2R is able to make this decision with a Precision of 80.45% and improves Recall nearly 8-fold from 2.55% without L2R to 19.90% with L2R.
2. Distinguishing various historical time period terms (e.g., *History, 16th Century* versus *History, 17th Century*). L2R is able to recommend these terms with a Precision of 86.41% and improves Recall from 0.79% without L2R, to 54.72% with L2R.
3. Improving Recall for Check Tags. Although our current machine learning approach is already at a high F-score level, L2R significantly improves Recall for these terms. We are exploring how to retain this Recall while maintaining the current high Precision for Check Tags.

### **Evaluation Plan**

We use the gold standard of NLM human indexing to determine MTI performance. MTI performance is summarized and presented to the Index Section on a monthly basis so that any performance issues may be addressed quickly. MTI is also externally evaluated within the International BioASQ Challenges. MTI has provided baselines for each year of the BioASQ Challenge where the organizers provide an unbiased evaluation of all of the participating systems also using NLM human indexing as the gold standard. One other metric we use to evaluate how well MTI is doing is to look at how often indexers actually reference the recommendations that MTI provides.

We started tracking MTI performance in 2007 and have made major improvements since that time as illustrated in Figure 5. MTI has performed well in all of the BioASQ Challenges consistently finishing in the top-tier, and a recent quote from one of the indexers nicely illustrates how useful they find the MTI recommendations: ". . . from our perspective, it's not so much that MTI is *STILL* useful to the task of indexing, it's that it is increasingly very useful to the task of indexing . . . there has been a real shift in perspective on MTI. Indexers used to view it as not helpful . . . now (most) view it as extremely helpful and overall very accurate".

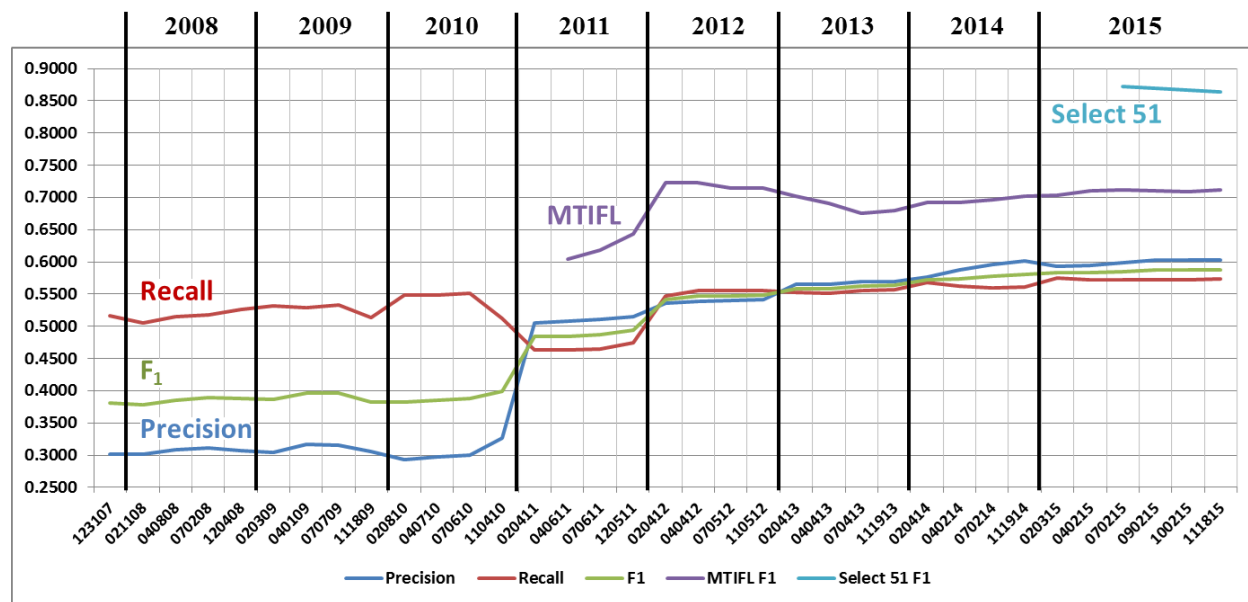


Figure 5 MTI Performance 2007 – Present (November 18, 2015)

In 2007, MTI statistics showed Precision of 0.3019, Recall of 0.5163, and  $F_1$  of 0.3810. In 2015, the MTI statistics show significant improvement in Precision and  $F_1$  with modest gains in Recall reflecting our focus on improving MTI Precision over the years: Precision of 0.6034 (+**0.3015**), Recall of 0.5730 (+**0.0567**), and  $F_1$  of 0.5878 (+**0.2068**). Figure 5 illustrates the performance changes of MTI between 2007 and 2015 using Precision, Recall, and  $F_1$  measures. Figure 5 also shows MTIFL  $F_1$  results between 2011 and 2015 and Select 51 Journal  $F_1$  results for 2015. It is clear from Figure 5 that journals added to the MTIFL program are some of the top performers with the  $F_1$  score (0.7113) dramatically higher than the overall MTI performance (0.5878). The Select set of 51 journals outperforms even the MTIFL journals by a wide margin with an  $F_1$  of 0.8642.

## Summary and Future Work

MTI, a project initiated to address practical NLM needs, developed into an exciting fully-fledged research project. As with many other NLM needs, there were no ready-to-use tools and even the first implementation of this “practical” tool required researching how to put together various sources of information, pioneering research at the time. Over the years, deeper understanding of the task allowed us to approach it as a multi-label classification problem: an information extraction and ranking task, a learning-to-rank problem, and an abstractive summarization task. These problems continue to be active areas of research, to which we continue contributing and at the same time apply the results of our research in practice.

Our next plans include improving over the achieved learning-to-rank results; employing more sophisticated methods for study subject extraction from full-text articles; exploring recently proposed supervised machine learning algorithms; and learning appropriate cutoff levels for the lists of suggestions.

As the Index Section considers more sophisticated indexing, we plan to research suggesting relation extraction to augment MH/SH suggestions. We will be reviewing related research as well as the previous Lister Hill Center work in this area [26] and expanding beyond it prioritizing the relations of interest to

the Index Section. For example, Figure 6 shows that in addition to current indexing nomenclature for the two MHs *Warfarin* and *Hemorrhage*, a clearer, more concise, and a more machine-friendly relation could be created using *Warfarin* **adverse effects** *Hemorrhage*. We are working with the Index Section to develop a pilot study investigating the relationship creation process and viability.

**Current Indexing Example:**

Warfarin/\*administration & dosage/adverse effects/blood  
Hemorrhage/chemically induced

**Possible Suggested Relation Extraction Example:**

**Warfarin** *adverse effects* **Hemorrhage**

Figure 6 Current Indexing vs Suggested Relation Extraction Example

We are also considering extracting more detailed information about study subjects, particularly for clinical trials participants from full text.



## References

**Note:** Author names of students who have worked on the MTI project are highlighted in **yellow**.

**Note:** Author names of indexers who have contributed to the MTI project are highlighted in **gray**.

1. Rajashekar TB, Croft BW. Combining automatic and manual index representations in probabilistic retrieval. *Journal of the American society for information science*. 1995;46(4):272-83.
2. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM Indexing Initiative. *Proc AMIA Symp 2000*;:17-21. PMID: 11079836
3. Mork JG, Jimeno-Yepes, A, Aronson AR. The NLM Medical Text Indexer System for Indexing Biomedical Literature. *BioASQ*. 2013.
4. Aronson AR. The MMI Ranking Function Whitepaper (1997). Available at <http://skr.nlm.nih.gov/papers/references/ranking.pdf>.
5. Aronson AR and Lang FM. (2010). An Overview of MetaMap: Historical Perspective and Recent Advances. *J Am Med Inform Assoc*. 2010 May 1;17(3):229-36. PMID: 20442139
6. Lin, J, & Wilbur, WJ (2007). PubMed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8(1), 423. PMID: 17971238
7. Jimeno-Yepes, A, Mork, JG, Demner-Fushman, D, Aronson, AR. Automatic algorithm selection for MeSH Heading indexing based on meta-learning. *International Symposium on Languages in Biology and Medicine*, Singapore, December, 2011.
8. Jimeno-Yepes, A, Mork JG, Demner-Fushman D, Aronson AR. Comparison and combination of several MeSH indexing approaches. *AMIA Annual Symposium Proceedings*. Vol. 2013. American Medical Informatics Association, 2013. PMID: 24551371
9. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond Synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies. *Proc AMIA Symp 1998*;:815-9. PMID: 9929332
10. Medical Text Indexer (MTI) Processing Flow Whitepaper. Available at [http://skr.nlm.nih.gov/resource/Medical\\_Text\\_Indexer\\_Processing\\_Flow.pdf](http://skr.nlm.nih.gov/resource/Medical_Text_Indexer_Processing_Flow.pdf).
11. Mork JG, Demner-Fushman D, Schmidt S, Aronson AR. Recent Enhancements to the NLM Medical Text Indexer. 2014 BioASQ Workshop at CLEF2014. Sheffield, England, September 2014.
12. Mork JG, Demner-Fushman D, Schmidt S, Aronson AR. Vocabulary Density Method for Customized Indexing of MEDLINE Journals. Poster AMIA 2014
13. Jimeno-Yepes A, Wilkowski B, Mork JG, Demner-Fushman D, Aronson AR. MeSH indexing: machine learning and lessons learned. *ACM SIGHIT International Health Informatics Symposium*, Miami, FL, USA, 2012.
14. Jimeno-Yepes A, Mork JG, Demner-Fushman D, Aronson AR. A one-size-fits-all indexing method does not exist: automatic selection based on meta-learning. *JCSE*, vol. 6, no. 2, pp.151-160, 2012.
15. Funk M, Reid C. Indexing Consistency in MEDLINE. *Bulletin of the Medical Library Association*. 71(2): 176-183; April 1983.
16. Névéol A, Mork JG, Aronson AR. Automatic Indexing of Specialized Documents: Using Generic vs. Domain-Specific Document Representations. *Proc BioNLP 2007 Workshop*, 183-92.
17. Névéol A, Shooshan SE, Humphrey SM, Rindflesch TC, Aronson AR. Multiple Approaches to Fine-Grained Indexing of the Biomedical Literature. *Proc Pacific Symposium on Biocomputing 2007*, 292-303. PMID: 17990500

18. Névél A, Shooshan SE, Mork JG, Aronson AR. Fine-Grained Indexing of the Biomedical Literature: MeSH Subheading Attachment for a MEDLINE Indexing Tool. *AMIA Annu Symp Proc.* 2007;;553-7. PMID: 18693897
19. Jimeno-Yepes A, Plaza L, Mork JG, Aronson AR, Díaz A. MeSH indexing based on automatically generated summaries. *BMC Bioinformatics.* 2013 Jun 26;14:208. doi: 10.1186/1471-2105-14-208. PMID: 23802936
20. Gay CW, Kayaalp M, Aronson AR. Semi-automatic indexing of full text biomedical articles. *AMIA Annu Symp Proc.* 2005;;271-5. PMID: 16779044
21. Demner-Fushman D, Mork JG. Extracting Characteristics of the Study Subjects from Full-Text Articles. *AMIA 2015.*
22. Liu K, Wu J, Peng S, Zhai C, Zhu S. “The Fudan-UIUC participation in the BioASQ Challenge Task 2a: The Antinomyra system”. *BioASQ 2014.*
23. Mao Y, Lu Z. NCBI at the 2013 BioASQ challenge task: Learning to rank for automatic MeSH indexing. *BioASQ.* 2013.
24. Tsoumakas G, Laliotis M, Markantonatos N, Vlahavas I. Large-Scale Semantic Indexing of Biomedical Publications. *BioASQ 2013.*
25. X. Liu K, Peng S, Wu J, Zhai C, Mamitsuka H, Zhu S. MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics*, 31(12), pp.i339-i347. 2015. PMID: 26072501
26. Winnenburger R, Sorbello A, Bodenreider O. Exploring adverse drug events at the class level. *J Biomed Semantics.* 2015 May 1;6:18. doi: 10.1186/s13326-015-0017-1. eCollection 2015. PMID: 25937884