Assessment of MTIFL Indexing Evaluation Methods

Project Leader

Susan Roy, NLM Associate Fellow 2011-2012

Project Sponsors

Rebecca Stanger, Unit Head, Index Section

Dr. Olga Printseva, Unit Head, Index Section

Spring/Summer 2012

Table of Contents

Abstract

OBJECTIVE: The purpose of this project was to assess methodologies which the Index Section of The National Library of Medicine® (NLM) could use for continuous evaluation of the Medical Text Indexer First Line (MTIFL), a semi-automated MEDLINE indexing system.

METHODS: To determine a potential method for indexing evaluation, one quantitative and two qualitative approaches were assessed. For the quantitative approach, statistics provided by the Lister Hill National Center for Biomedical Communications (LHC) to monitor the precision (P), recall (R) and F-scores of Medical Subject Headings® (MeSH) suggested by MTIFL were evaluated. One of the qualitative approaches involved NLM Senior Indexers analyzing the MeSH terms indexed via the MTIFL indexing process. The second qualitative approach involved NLM indexers rating MEDLINE citations for the appropriateness of MeSH terms. The three approaches were tested then analyzed based on the resulting criteria of evaluative data, feasibility, and efficiency for continual indexing evaluation.

RESULTS: The LHC statistics for the P, R and F-scores allow for efficient quantitative analysis of MeSH terms added or deleted during MTIFL indexing but these data do not provide an adequate assessment for the quality of final indexing. The Senior Indexer Analysis yielded in-depth details about the quality of indexing, including the reasons why MeSH terms are added and/or deleted during the MTIFL indexing process as well as details about critical MeSH terms needed for indexing of articles. But the approach was found to be too labor-intensive and not a feasible on-going option. Finally, the Indexer Rating Survey revealed final indexing quality in an easier-to-use methodology.

CONCLUSIONS: This study investigated possible ways in which NLM could evaluate the MTIFL indexing. Small-scale studies involving the data collection of P, R and F-scores, a critical analysis of the indexing process, and ratings of MeSH terms applied on indexed articles were tested and evaluated as potential models. A combined approach utilizing measures from each of the three methods described is proposed for the evaluation of MTIFL indexing.

Introduction

Indexing at the National Library of Medicine® (NLM) involves an indexer who will scan an article and then assign subject terms as needed to reflect the content of the article. These terms used by the indexers are Medical Subject Headings® (MeSH), the controlled vocabulary created and curated by the NLM to describe topics that are represented in the biomedical article. Today MeSH contains over 26,000 subject headings (or descriptors) that are arranged in a hierarchy with the option of adding subheadings to narrow, or focus, on the topic. Supplementary concept records (SCR) can also be included to describe chemical substances and diseases. MeSH terms that are considered to represent a main idea or focus of an article are also referred to as IM (*Index Medicus*), starred, asterisked or Major Topic terms. The assignment of MeSH descriptors and subheadings is important for proper citation retrieval in MEDLINE®, the NLM bibliographic database of biomedical citations. Indexers tend to be experts in a particular field of biomedicine and/or life sciences in addition to being extensively trained in the principles of MEDLINE indexing.

The labor-intensive task of indexing, combined with the rapid increase in number of articles per issue and journals being selected for MEDLINE, has led to the adoption of the Indexing Initiative (II) (1, 2). The goal of the II is to determine methods where automation can assist in the indexing of articles. Index Section staff has worked with the Cognitive Science Branch of the Lister Hill National Center for Biomedical Communications (LHNCBC) to develop the Medical Text Indexer (MTI) to discover ways to promote automation assistance in indexing.

MTI is a software program that is used to assist in the indexing of biomedical literature by suggesting MeSH terms (1). MTI software uses two methods to index articles both of which show optimal results when text mining a title (TI) and abstract (AB). The first method utilizes MetaMap Indexing, which ranks concepts by extracting the noun phrases from the text and then computes Unified Medical Language System ® (UMLS) concepts (3). The second method utilizes PubMed Related Citations, an algorithm which finds other citations statistically related to the citation being indexed to find similar MeSH terms (4). Once UMLS concepts and related citation concepts are located, MTI restricts to MeSH terms by using synonyms through associated expressions and inter-concept relationships. Finally, MeSH terms are ranked to give suggested index terminology (1, 4, 5).

Since 2002 MTI recommendations are viewed on the NLM Data Creation and Maintenance System (DCMS), a Web-based interface used to index the MEDLINE citations. On the DCMS indexers have the option to consult MTI MeSH term suggestions or not. Previous research compared MTI recommendations to human indexing, and found that the MTI does relatively well and correctly recommends MeSH terms 51.07% of the time (4). Most interesting was the finding that when MTI indexes with both a TI and AB, significantly better results are shown than when indexed from TI alone (53.79% and 29.91%, respectively) (4). There has been steady improvement with MTI suggestions. As of 2011, MTI provides recommendations for over 96% of the total number of citations indexed. Citations from 2011 shows MTI had an overall F-

measure of 49.40% (Recall: 47.40%, Precision: 51.57%) whereas articles that had both a TI and AB, the F-measure was 50.60% (Recall: 50.11%, Precision: 51.10%) (2).

These results led to journals being selected for MTI First Line (MTIFL) indexing. MTIFL utilizes the suggested MeSH terms from MTI and algorithms optimized for precision to index MeSH terms. Because of this optimization, MTIFL will first start the indexing of the article and an indexer will complete the indexing by adding and deleting MeSH terms as appropriate when consulting the full-text of the article. The current goal is to continue to pursue the use of MTIFL because implementation has contributed to a more efficient indexing workflow, saving time and money. In 2011 a total of 23 journals was selected for MTIFL indexing and in 2012 an additional 22 were promoted to MTIFL status. Because of the increased citation indexing workload on the indexers and the promise of MTIFL, it has become necessary to determine methods for evaluating indexing, especially that of MTIFL.

Traditional measures of evaluation for bioinformatic applications have utilized precision and recall. These scores attempt to measure the effectiveness of the retrieval system. Essentially, precision is how well the program retrieved only the most relevant documents (i.e., relevant and retrieved documents divided by all retrieved documents). Recall is how well a program retrieves all of the documents (i.e., relevant and retrieved documents divided by all relevant documents). Although these quantitative measures have been widely used in evaluation methodologies, these measures might not always provide an accurate in-depth quality description. For example, precision and recall do not actually evaluate the quality of indexing, and reliance solely on statistical and quantitative methods might not provide the best assessment for MEDLINE indexing quality. Here we hypothesize that a combined quantitative and qualitative approach might be better for the evaluation of MTIFL indexing. Therefore, the purpose of this project was to determine a model that the NLM Index Section could use for simple and continuous evaluation of MTIFL indexing. A discussion about lessons learned, ease of effectiveness, feasibility and efficiency in the tested approaches, along with possible suggestions for future approaches, are described.

Methods

To determine a method of MTIFL indexing evaluation, one quantitative and two qualitative approaches were assessed. A description of how the data were tested will be presented followed by an evaluation of the three approaches.

Selection of Journals for Pilot Studies

At the start of the project, the project sponsors selected four journals from the original 14 MTIFL indexed journals (these four selected journals had the most completed issues available in MEDLINE). These four journals were:

- *Archives of Microbiology (Arch Microbiol)*
- *Canadian Journal of Microbiology (Can J Microbiol)*
- *ISME J (International Society for Microbial Ecology Journal) (ISME J)*
- *Journal of Applied Microbiology (J Appl Microbiol)*

For each of the four journals, three issues from 2009 (indexed by the traditional human method) and the corresponding three issues from 2011 (indexed by the MTIFL method) were selected. Using the Single Citation Search function in PubMed all six issues for each of the four journals were searched using the following example search query:

"Archives of microbiology"[Jour] AND 191[volume] AND 8[issue] AND 2009[pdat]

The MEDLINE display file was exported to Excel™ and the PubMed Unique Identifier (PMID), article title (TI), MeSH Heading (MH) and Registry Number (RN) elements were retained and all other metatagged data were removed. The number and averages of MeSH terms, IM terms and SCRs (the RN element) for each issue were determined and then normalized by dividing by the number of total articles. Following these analyses, we were able to make a decision to focus on two journals (*J Appl Microbiol* and *ISME J*) in the subsequent pilot studies on evaluation methods.

Lister Hill Center Statistics

With the assistance of Lister Hill Center (LHC) staff, precision (P), recall (R) and F-scores (F) of MTIFL-indexed MeSH terms were generated. Precision was calculated as the percent of retained MTIFL-indexed MeSH (i.e., the MeSH terms kept by the indexers that were originally indexed by MTIFL) terms divided by the number of original MTIFL-indexed MeSH terms. Recall was calculated as the percent of retained MeSH terms divided by the final number of indexed MeSH terms. The F-score was calculated by combining recall and precision into a single number, which is a measure typically used to verify the accuracy and performance of information retrieval programs. The F-score can be considered a weighted average of P and R because both measures are used in the calculation (see Appendix B for additional information about MTI statistics).

The P, R and F data were generated for the three 2011 MTIFL-indexed issues for both journals, *J Appl Microbiol* and *ISME J*. Data for an additional smaller subset of 12 semi-randomly selected articles were examined in a separate analysis. The 12 semi-randomly selected articles were chosen by fixing the number of articles randomly chosen per journal and issue (i.e., two articles from each of the three issues were chosen, for a total of six articles from *J Appl Microbiol* and six articles from *ISME J*). In addition to the P, R and F data, LHC-generated data for reasons why MTIFL-indexed MeSH terms were added and/or deleted.

Senior Indexer Critical Analysis

One qualitative approach that was assessed was an in-depth analysis of final indexed MeSH terms. Two expert NLM Senior Indexers examined the same 12 MITFL-indexed articles (six *J Appl Microbiol* and six *ISME J*) used for the LHC analysis based on the following measures: original MeSH terms indexed by MTIFL, reasons why indexers deleted or added MeSH terms, and an analysis of IM terms. These were the same 12 articles from the semi-random selection described above. Particular attention was paid to MeSH terms that were deemed critical to the indexing of the article. Rational for the deleted and added MeSH terms was categorized for quantitative analysis.

Indexer Ratings

The second qualitative approach tested utilized a survey and rating method. Four NLM indexers were given 24 articles (TI, AB and corresponding indexed MeSH terms only) and asked to rate the appropriateness of indexed MeSH and IM terms. 12 articles from *J Appl Microbiol* and 12 articles from *ISME J* were semi-randomly selected. For both journals, six 2009 articles (indexed by humans) and six 2011 articles (indexed via the MTIFL method) were chosen at random (the 12 MTIFL-indexed articles were the same 12 that were used in the previous two assessment studies). All indexers were given the same articles in a randomized order and were blinded to the method of indexing. The indexers were given the instructions, questions, and Likert scale for ratings (see Figure 1). The data were collected and analyzed to determine if indexers were satisfied with the final indexed MeSH and IM terms.

Please look at the Title and Abstract as you would if you were indexing. Then please look at the MeSH terms and the IM terms that have been applied to the citation. Then answer the questions below for each citation.

1 – Are you satisfied with the MeSH terms that have been applied to this article?

2 – Are you satisfied with the IM terms that have been applied to this article?

Strongly Agree	Somewhat Agree	Undecided	Somewhat Disagree	Strongly Disagree
1	2	3	4	5

Figure 1. Instructions, questions and Likert scale used in the Indexer Survey Ratings.

Statistics

Where appropriate, a Student's two-tailed, unpaired t-test was completed to statistically compare the means of groups. An alpha level of 0.05 was used in all statistical tests for significance.

Results

Selection of Journals for Pilot Studies

To determine which journals and issues to use in our pilot studies, we analyzed the number of articles, MeSH and IM terms in four journals (see Appendix A and Table 1). To compare the differences between the 2009 (human) and 2011 (MTIFL) indexing, t-test analyses were conducted (see Table 2). Only *J Appl Microbiol* had significant differences in the number of articles, MeSH and IM terms indexed and *ISME J* had significant different number of articles form 2009 compared to 2011. Neither of the other two journals (*Arch Microbiol* or *Can J Microbiol*) showed any observable differences in measures between the 2009 to 2011 issues.

Indexing Method	Avg Articles	Avg MeSH	Avg IM	Avg SCR
Arch Microbiol				
Human	6.33	74.33	27.33	31.33
MTIFL	7.33	80.67	28.33	34.67
Can J Microbiol				
Human	13.00	153.67	56.67	45.67
MTIFL	10.00	116.33	41.00	37.67
ISME J				
Human	9.67	112.67	41.33	28.00
MTIFL	14.00	129.67	54.33	35.67
J Appl Microbiol				
Human	36.67	462.67	163.00	138.67
MTIFL	25.33	303.00	109.67	97.33

Table 1. Average number of articles, MeSH, IM and SCR from three human-indexed (2009) or three MTIFL-indexed (2011) issues of *Arch Microbiol*, *Can J Microbiol*, *ISME J* and *J Appl Microbiol*.

Student's two-tailed, unpaired t-tests					
Journal	**Measure**	***t***	**df**	***p***	**significance**
Arch Microbiol	Articles	1.3416	4	0.2508	ns
	MeSH	0.4952	4	0.6464	ns
	IM	0.2458	4	0.818	ns
Can J Microbiol	Articles	1.1078	4	0.3301	ns
	MeSH	1.0217	4	0.3647	ns
	IM	1.2703	4	0.2728	ns
ISME J	Articles	3.6056	4	0.0226	Yes
	MeSH	0.8339	4	0.4512	ns
	IM	1.5192	4	0.2033	ns
J Appl Microbiol	Articles	12.0208	4	0.0003	Yes
	MeSH	6.6197	4	0.0027	Yes
	IM	7.7703	4	0.0015	Yes

Table 2. Article, MeSH and IM term t-test data for the four analyzed journals. Only *J Appl Microbiol* and *ISME J* showed significant observable differences in measures. (ns = not significant)

Because journals had different number of articles per issue, the number of MeSH and IM terms were normalized by dividing them by the total number of articles from all issues analyzed (see Figure 2). Following these analyses, it was determined that *ISME J* and *J Appl Microbiol* might be suitable candidates to use in the subsequent pilot tests to determine a potential method for the evaluation of indexing because they displayed the greatest differences in the number of MeSH and IM terms. That is, it was reasoned that these would be suitable candidates to use in our pilot studies because we might see differences in MeSH and IM term quality in addition to differences in quantity.

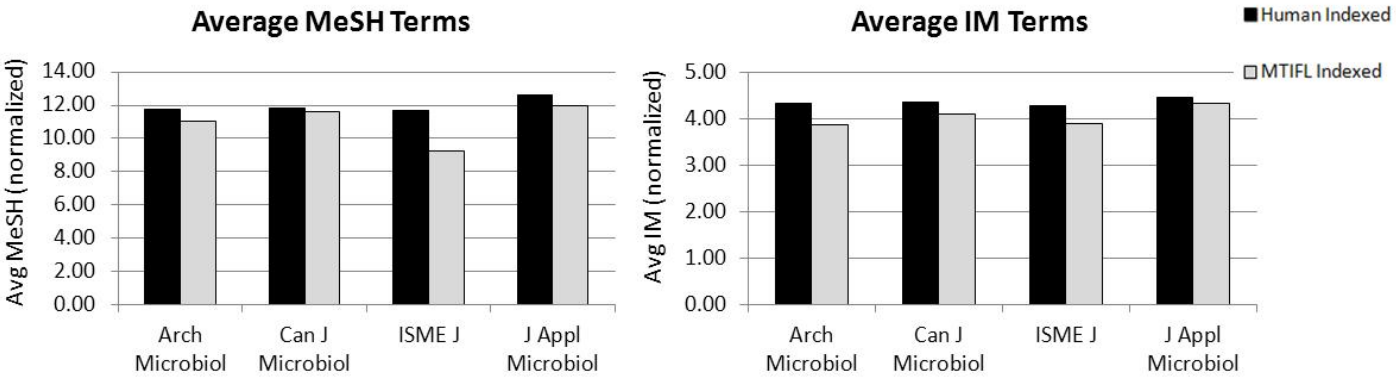

Figure 2. Average number of MeSH (left) and IM (right) terms normalized by total number of articles for all four journals. The dark bar shows averages for the analyzed three 2009 (human-indexed) issues and the light bar for three 2011 (MTIFL-indexed) issues.

Lister Hill Center Statistics

To evaluate the usability of LHC statistics as a method for MTIFL indexing method evaluation, the precision, recall and F-scores for the three 2011 issues of *J Appl Microbiol* and *ISME J* were generated. The three issues of *J Appl Microbiol* consisted of 76 articles. These 76 articles had a total of 905 final indexed MeSH terms. MTIFL originally indexed 776 MeSH terms, 248 (31.96%) were removed, 377 (41.66%) were added and 528 were retained by the indexers. For these 76 *J Appl Microbiol* articles, a recall of 58.34%, precision of 68.04%, and an F-score of 62.82% was calculated. The three issues of *ISME J* consisted of 42 *ISME J* articles. These 42 articles had a total of 387 final indexed MeSH terms where MTIFL originally indexed 393 MeSH terms, 116 (29.52%) were removed, 110 (28.42%) were added and 277 were retained by the indexers. These data gave a recall of 71.58%, precision of 70.48%, and F-score of 71.03% (see Table 3 and Figure 3).

	***J Appl Microbiol* (76 articles)**		***ISME J* (42 articles)**	
	# MeSH	**Percent**	**# MeSH**	**Percent**
MTIFL Indexed	776		393	
Final Indexed	905		387	
MeSH Terms Added	377	41.66%	110	28.42%
MeSH Terms Removed	248	31.96%	116	29.52%
MeSH Terms Retained	528		277	
Recall		58.34%		71.58%
Precision		68.04%		70.48%
F-Score		62.82%		71.03%

Table 3. LHC-generated MeSH term counts for MTIFL-indexed articles.

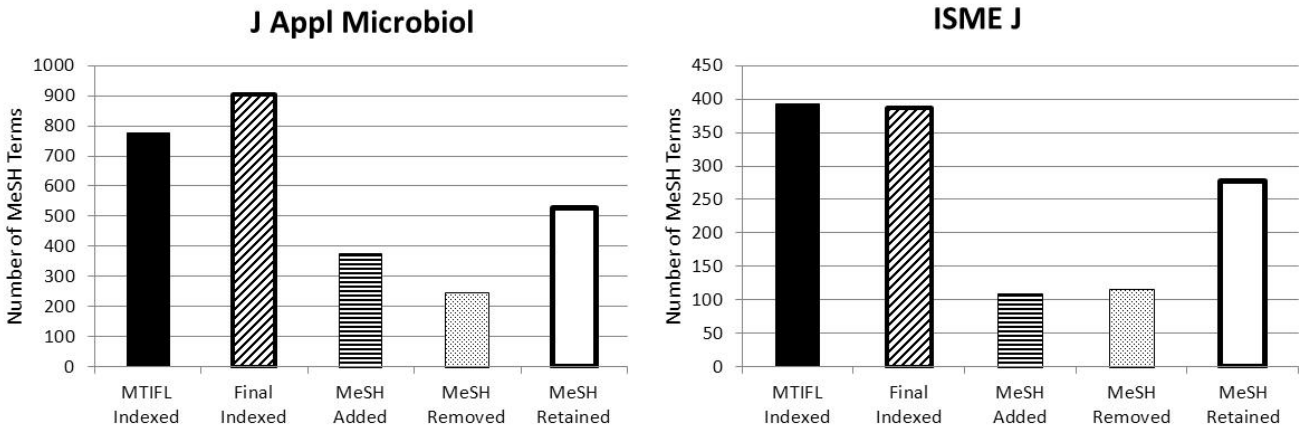

Figure 3. Displays the total number of indexed MeSH terms during the MTIFL indexing process for 72 *J Appl Microbiol* (left) articles and 42 *ISME J* (right) articles.

For subsequent analyses, a smaller sample of articles from these issues was selected and analyzed for ease of managing the evaluation methodology assessment. Six articles from *J Appl Microbiol* had a total of 73 final indexed MeSH terms. MTIFL originally indexed 79 MeSH terms, 35 (44.30%) were removed, 29 (39.73%) were added and 44 were retained by indexers. These 6 J Appl Microbiol articles had a recall of 60.27%, precision of 55.70%, and F-score of 57.89%. For the six *ISME J* articles, 71 MeSH terms were originally indexed via MTIFL, 21 (29.58%) were removed, 10 (16.67%) were added and 50 were retained for a final indexing of 60 MeSH terms for a recall of 83.33%, precision 70.42% and F-score 76.34% (see Table 4 and Figure 4).

	***J Appl Microbiol* (6 articles)**		***ISME J* (6 articles)**	
	# MeSH	**Percent**	**# MeSH**	**Percent**
MTIFL Indexed	79		71	
Final Indexed	73		60	
MeSH Terms Added	29	39.73%	10	16.67%
MeSH Terms Removed	35	44.30%	21	29.58%
MeSH Terms Retained	44		50	
Recall		60.27%		83.33%
Precision		55.70%		70.42%
F-Score		57.89%		76.34%

Table 4. LHC-generated MeSH term counts for a smaller subset of MTIFL-indexed articles. The 6 articles analyzed for *J Appl Microbiol* had an F-score of 57.89% and the 6 *ISME J* articles had an F-score of 76.34%.

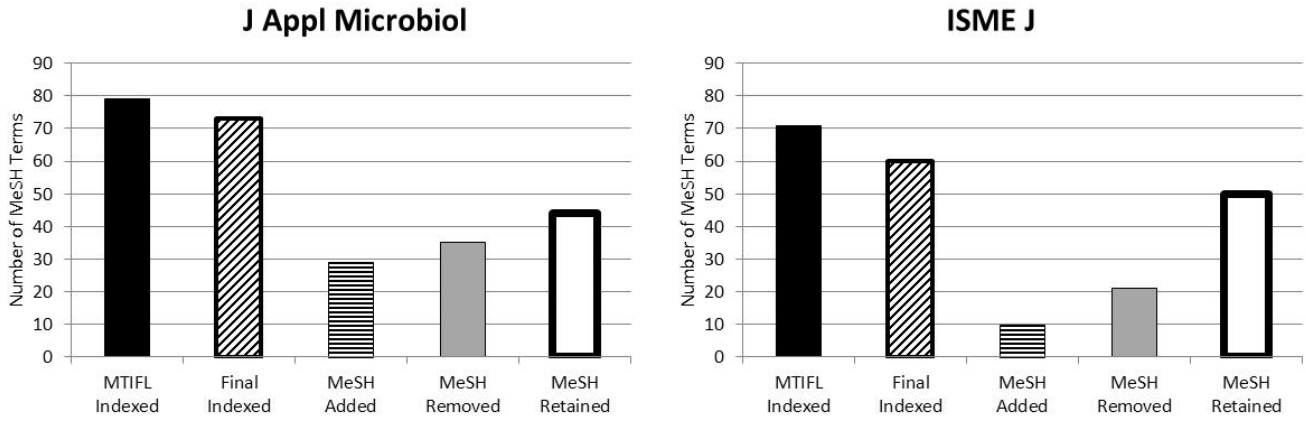

Figure 4. Displays the total number of MeSH terms during the MTIFL indexing process for the 6 *J Appl Microbiol* (left) and 6 *ISME J* (right) articles.

In addition to MeSH term counts, P, R and F-data, LHC can also potentially provide details about MTIFL indexing by comparing the indexer deleted and added MeSH terms. That is, by comparing MTIFL deleted MeSH terms to indexer added MeSH terms, LHC statistics could provide rationale on why the indexer chose to delete and add a different MeSH term (see Table 5 and Figure 5). The definitions for the measures provided by LHC are as follows:

- **Not Matched, More Specific** – Deleted MeSH term was found to be further down the same branch as an unmatched Indexed MH
- **Not Matched, More General** – Deleted MeSH term was found to be further up the MeSH Tree than an unmatched Indexed MH
- **Matched, More Specific** – Deleted MeSH term was found to be further down the MeSH Tree than a matched Indexed MH
- **Matched, More General** – Deleted MeSH term was found to be further up the MeSH Tree than a matched Indexed MH
- **Not Matched, Same Category** – Deleted MeSH term was found to be in the same top level MeSH Category as an unmatched Indexed MH
- **Matched, Same Category** – Deleted MeSH term was found to be in the same top level MeSH Category as a matched Indexed MH
- **Out of the Ballpark** – Deleted MeSH term was not found to match any Indexed MH MeSH Category
- **CheckTags Not Reviewed** – Deleted MeSH term that are considered CheckTags were not reviewed
- **One Level Down** - MTI term down one level in MeSH Tree from Human-indexed IM term
- **One Level Up** - MTI term up one level in MeSH Tree from Human-indexed IM term
- **Same Tree Branch** - MTI Recommendation is in the same MeSH Tree branch as Human-indexed IM term
- **IM Missed Completely** - None of the MTI Recommendations were a close match to a missed IM Term

		J Appl Microbiol		ISME J	
Reason	**Tree Level**	**Count**	**Percent (%)**	**Count**	**Percent (%)**
Not Matched, More Specific	1	0	0.00	1	4.76
	2	0	0.00	0	0.00
	3+	0	0.00	0	0.00
	Total	0	0.00	1	4.76
Not Matched, More General	1	3	8.57	0	0.00
	2	1	2.86	0	0.00
	3+	0	0.00	0	0.00
	Total	4	11.43	0	0.00
Matched, More Specific	1	1	2.86	3	14.29
	2	0	0.00	1	4.76
	3+	6	17.14	0	0.00
	Total	7	20.00	4	19.05
Matched, More General	1	5	14.29	1	4.76
	2	1	2.86	0	0.00
	3+	1	2.86	1	4.76
	Total	7	20.00	2	9.52
Not Matched, Same Category		6	17.14	7	33.33
Matched, Same Category		2	5.71	6	28.57
Out of the Ballpark		3	8.57	1	4.76
CheckTags Not Reviewed		6	17.14	0	0.00
Total MeSH Deleted		**35**		**21**	

Table 5. LHC-generated evaluation data for the MTIFL indexed MeSH terms changed by indexers.

The reasons why MeSH terms were deleted were varied making it difficult to generalize from this small sample size. *Matched, More Specific* and *Not Matched, Same Category* were more prominent, suggesting MeSH terms were deleted and more specific MeSH terms were added because the original MTIFL-indexed terms were too general (see Table 5 and Figure 5).

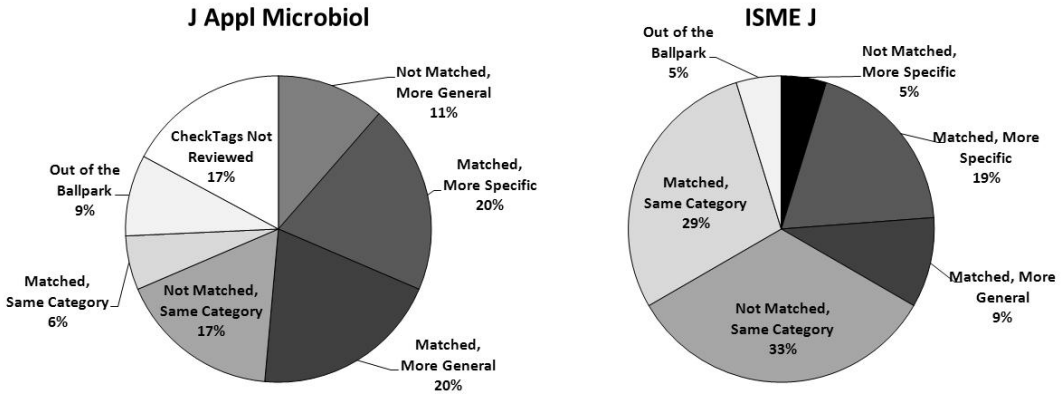

Figure 5. LHC-generated evaluation data (in percent) for MTIFL indexed MeSH terms changed by indexers.

LHC also generated evaluative data for the reasons why IM terms were deleted by comparing human added to deleted MTIFL-indexed IM terms (see Figure 6). Because of our small sample size, not many IM terms could be evaluated (six from *J Appl Microbiol* and one from *ISME J*). For *J Appl Microbiol*, 67% (or 4 out of the 6) IM terms fit the category of *Out of the Ballpark*. The other two IM terms for *J Appl Microbiol* were found to be either in the *Same Tree* or *One Level Up* when compared to the added IM term.

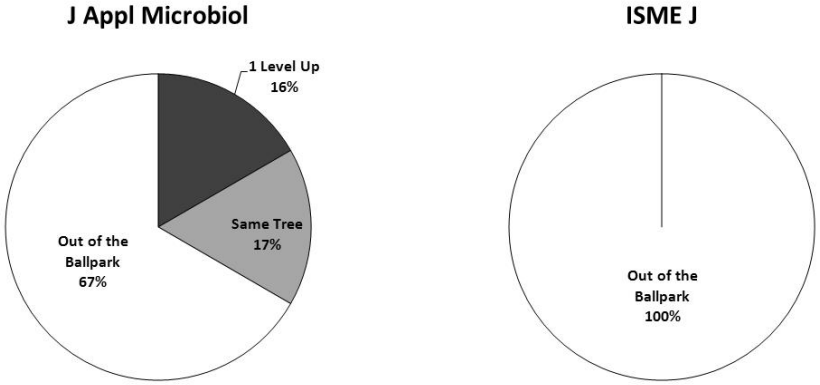

Figure 6. LHC-generated evaluation data (in percent) for IM terms changed by indexers.

Senior Indexer Analysis

Two NLM Senior Indexers evaluated indexing on 12 MTIFL-indexed articles. First, the number of MeSH terms indexed by MTIFL, MeSH terms deleted and added by indexers, IM terms, critical MeSH terms and final number of indexed terms were calculated and averaged (see Table 6 and Figure 7). A Student's two-tailed, unpaired t-test determined there were no significant differences between the number of MeSH terms indexed by MTIFL compared to the final number of MeSH terms indexed after completion by human indexer (*J Appl Microbiol* $t(10) = 0.2007$, $p=0.8449$, and *ISME J* $t(10) = 0.7187$, $p=0.4888$). Although MeSH term counts were not exactly similar, the data correspond to those statistics from LHC, as reported above. The discrepancies are most likely due to an experimenter counting error.

	MTIFL-Indexed MeSH	Deleted MeSH	Added MeSH	IM Terms	Critical MeSH	Final Indexing
J Appl Microbiol						
Sum	76	30	27	6	25	73
Average	12.67	5.00	4.50	1.00	4.17	12.17
ISME J						
Sum	70	21	11	2	8	60
Average	11.67	3.50	1.83	0.33	1.33	10.00

Table 6. Total and average number of MTIFL-indexed MeSH terms, MeSH terms added and deleted by indexers, IM terms, critical MeSH terms and final indexed MeSH terms for six *J Appl Microbiol* and 6 *ISME J* articles, as determined by the Senior Indexer Analysis.

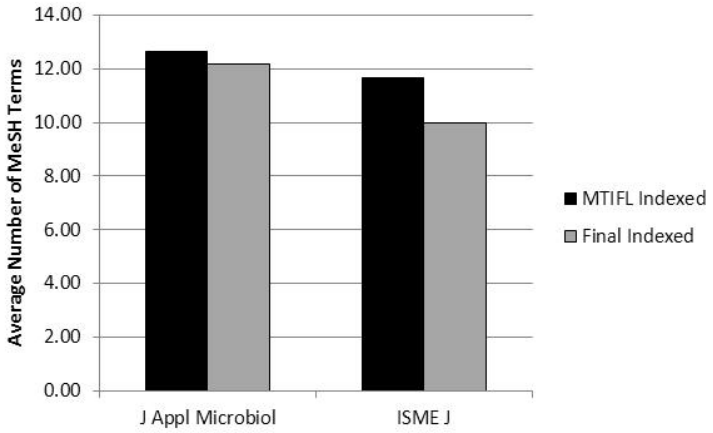

Figure 7. Average number of MeSH terms for six *J Appl Microbiol* and six *ISME J* articles. MTIFL-indexed MeSH term counts (dark bar) and final number of MeSH terms indexed (light bar) are displayed.

Next, the Senior Indexers evaluated MeSH terms that were deleted by the indexers. For the *J Appl Microbiol* articles, MTIFL originally indexed 76 MeSH terms and 30 were deleted by indexers (average 12.67 indexed, 5 deleted per article). These results were statistically significant, suggesting the indexers made significant changes to MTIFL indexing by deleting terms ($t(10) = 3.2658$, $p=0.0085$). Similarly for *ISME J* articles, MTIFL originally indexed 70 MeSH terms and 21 were deleted by indexers (average 11.67 indexed and 3.50 deleted per article) and these were found to be significantly different ($t(10) = 3.4426$, $p=0.0063$). Again, these results were similar to those described in the LHC statistics.

To evaluate why MTIFL-indexed MeSH terms were deleted, the Senior Indexers grouped reasons deleted into six categories:

- *Incorrect* – Wrong MeSH term
- *Too Tangential* – Inappropriate
- *Too General* – Higher up in the MeSH tree
- *Too Specific* – Lower down in the MeSH tree
- *3rd Tier Term* – Not the point, but can be used
- *Incorrectly Deleted* – Should not have been deleted

Figure 8 shows the categorical representation for the deleted terms in a percentage. For *J Appl Microbiol* the deleted terms primarily fell into the categories of *Too Tangential* (32%), or *Too General* (32%). For *ISME J*, deleted terms primarily fell into the categories of *Incorrectly Deleted* (29%), or *Too General* (19%). All other categories were also near equal in representation and with this small sample size only generalizations can be made about why MeSH terms were deleted.

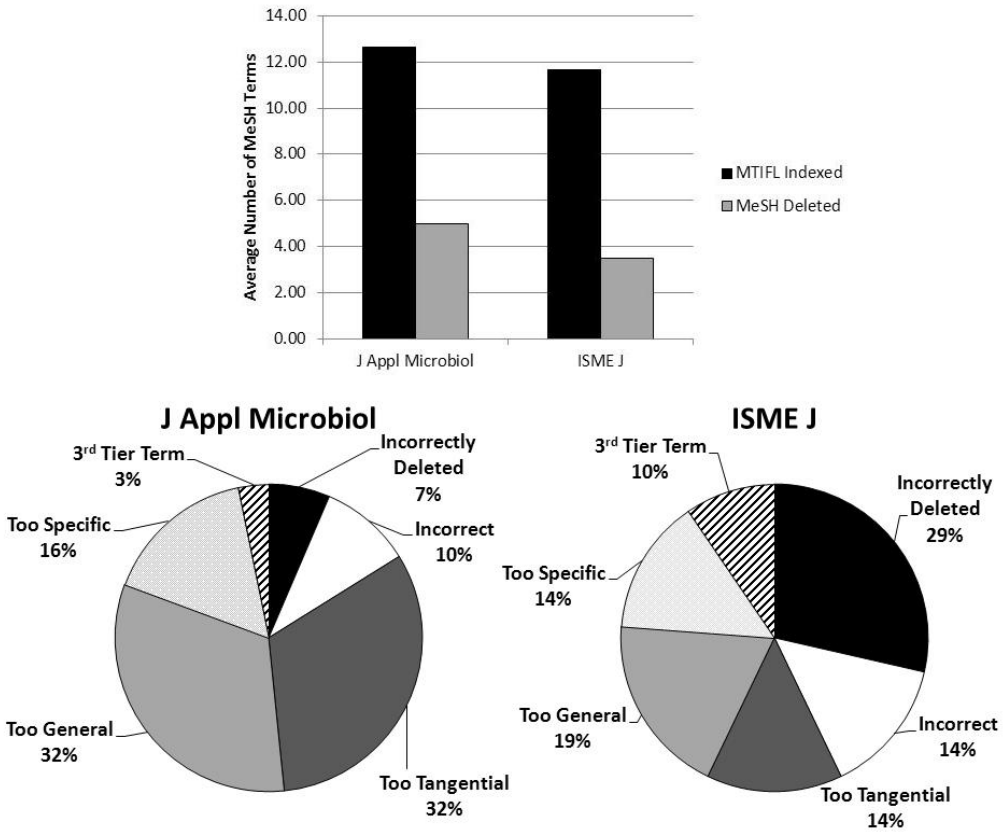

Figure 8. Top panel displays the average number of MeSH terms indexed by MTIFL (dark bars) and deleted by indexers (light bars) and the bottom panel displays categorical representation on why the MeSH terms were deleted by the indexers.

Next, the Senior Indexers evaluated MeSH terms that were added by the indexers. For the *J Appl Microbiol* articles 27 (average 4.5 per article) and for *ISME J* articles 11 (average 1.83 per article) MeSH terms were added by indexers. The number of MeSH terms added during the indexing process was not found to be a significant number when compared to the final number of MeSH terms; *J Appl Microbiol*, $t(10) = 0.2806$, $p=0.7847$; and *ISME J*, $t(10)=0.9583$, $p=0.3605$. However of the 27 MeSH terms added for *J Appl Microbiol*, 25 (4.17 on average, or 93%), were deemed critical to the indexing of the article and for *ISME J*, 8 of the 11 added terms (1.33 on average, or 73%), were deemed to be critical (see Figure 9, top panel). These results show that even though the number of MeSH terms added was not significant in comparison to the total number of terms indexed, a significant number of the terms added were deemed to be critical to the indexing of the article. These results provide important evidence that the indexers are a vital component of the MTIFL-method of indexing, performing a necessary step in the indexing process.

To evaluate why indexers added additional MeSH terms, the Senior Indexers grouped reasons added into five categories:

- *Point of Article* – Necessary for indexing
- *Title Concept* – Necessary for indexing
- *Indexing Policy* – Coordination and subheadings
- *Can Be Indexed* – Not the point but can be used
- *Found in Full-Text* – Found in other than title and abstract

Figure 9 (bottom panel) shows the categorical representation for the deleted terms in a percentage. For *J Appl Microbiol* most (41%) of the MeSH terms were added because those concepts were *Found In Full-Text* of the article and another 35% of the MeSH terms were added because they were the *Point of Article*. For the *ISME J* articles evaluated, MeSH terms were found to be added because they were deemed as a *Title Concept* (75%) or because they were *Found in the Full-Text* (25%) of the article. Again, these results provide evidence that indexers serve a critical role in indexing of MTIFL-indexed articles. Many of the added MeSH terms come from scanning the entire full-text of the article.

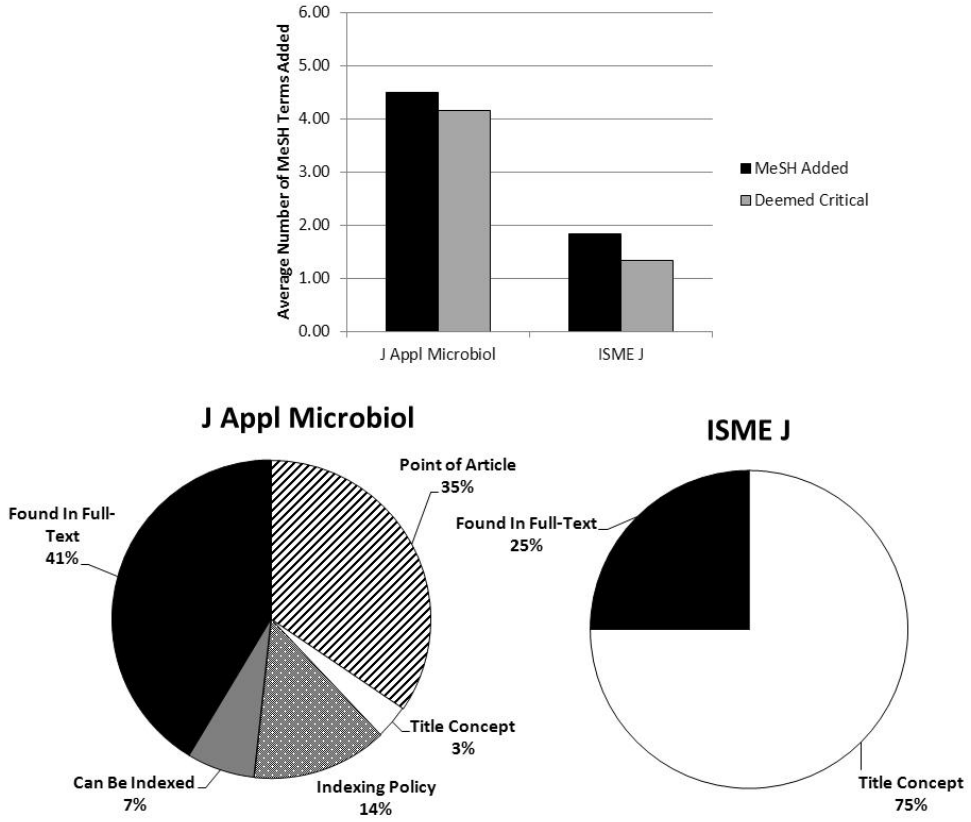

Figure 9. Top panel displays the average number of MeSH terms added by indexers (dark bar) and deemed critical (light bars) by Senior Indexers and the bottom panels displays the categorical representation on why the MeSH terms were added by the indexers.

Indexer Ratings

The final qualitative approach assessed looked to determine if a quick rating of MeSH and IM terms could be used to evaluate indexing. Four NLM indexers were given a semi-random set of articles and asked to rate how satisfied they were with the MeSH and IM terms indexed, ranging from *Strongly Agree* (1) to *Strongly Disagree* (5) (see Figure 1). For *J Appl Microbiol*, the MeSH and IM terms were rated slightly better (1 - *Strongly Agree* to 2 - *Somewhat Agree*) for MTIFL-indexed articles (1.58 and 1.54, respectively) compared to the human-indexed articles (2.29 and 2.38). These results were found to be significantly different; $t(46) = 2.1704$, $p=0.0352$, for MeSH term ratings, and $t(46) = 2.4949$, $p=0.0163$, for IM ratings (see Table 8).

For *ISME J*, the MeSH and IM term ratings for the two different modes of indexing were not found to be as different, when rated by indexers. MTIFL indexing was rated 1.96 (MeSH) and 1.96 (IM), whereas human indexing was rated 1.88 (MeSH) and 1.79 (IM). No significant differences in ratings were found for MeSH term ratings $t(46) = 0.2838$, $p=0.7778$; or IM ratings $t(46) = 0.5599$, $p=0.5783$ (see Table 8). These results show that the MeSH and IM terms indexed by both the human indexers and the MTIFL-assisted methods were rated satisfactory in indexing the articles.

Journal	Question	Index Method	Avg Rating	*t*	df	*p*	sig
J Appl Microbiol	1 - MeSH	Human	2.29	2.1704	46	0.0352	Yes
	1 - MeSH	MTIFL	1.58				
	2 – IM	Human	2.38	2.4949	46	0.0163	Yes
	2 - IM	MTIFL	1.54				
ISME J	1 - MeSH	Human	1.88	0.2838	46	0.7778	ns
	1 - MeSH	MTIFL	1.96				
	2 - IM	Human	1.79	0.5599	46	0.5783	ns
	2 - IM	MTIFL	1.96				

Table 8. Average MeSH and IM term ratings and t-test results for the 12 J Appl Microbiol (6 human and 6 MTIFL-indexed) and 12 ISMEJ (6 human and 6 MTIFL-indexed) articles, as rated by four NLM Indexers.

Figure 10 displays the percent distribution of ratings for both journals, and for both methods of indexing. Overall, the indexers rated the MeSH and IM terms (for both human and MTIFL-assisted indexing methods) as *Strongly Agree* to *Somewhat Agree* more often. Percentage-wise, both methods of indexing were rated on the positive side of the continuum, and represented greater than 60% of the ratings. Interestingly, for the *J Appl Microbiol* human0indexed articles, indexers *Strongly Disagreed* with MeSH and IM term indexing approximately 4% of the time. For the *ISMEJ* MTIFL-indexed articles indexers *Strongly Disagreed* with MeSH and IM term indexing about 4% of the time. However, because both methods, on all measures, rated more on the positive side of the continuum, we can say the indexing of these articles, regardless of the method of indexing, was deemed as sufficient. In fact, for the MTIFL and human methods of indexing, *ISME J* indexing was rated as more satisfactory approximately 80% of the time. For the MTIFL method on *J Appl Microbiol*, the indexing was rated over 90% as positive.

After completion of the Indexer Rating Survey it was discovered that the SCR terms were accidently omitted from the articles needing to be rated. The SCR terms were then pulled for the articles (6 of the 24 articles had SCR terms) and the indexers were asked if they would change their original ratings based on the new information about indexed SCR terms. Few changes were made to the original ratings for the indexed MeSH and/or IM terms, therefore, the final results still stand as presented above.

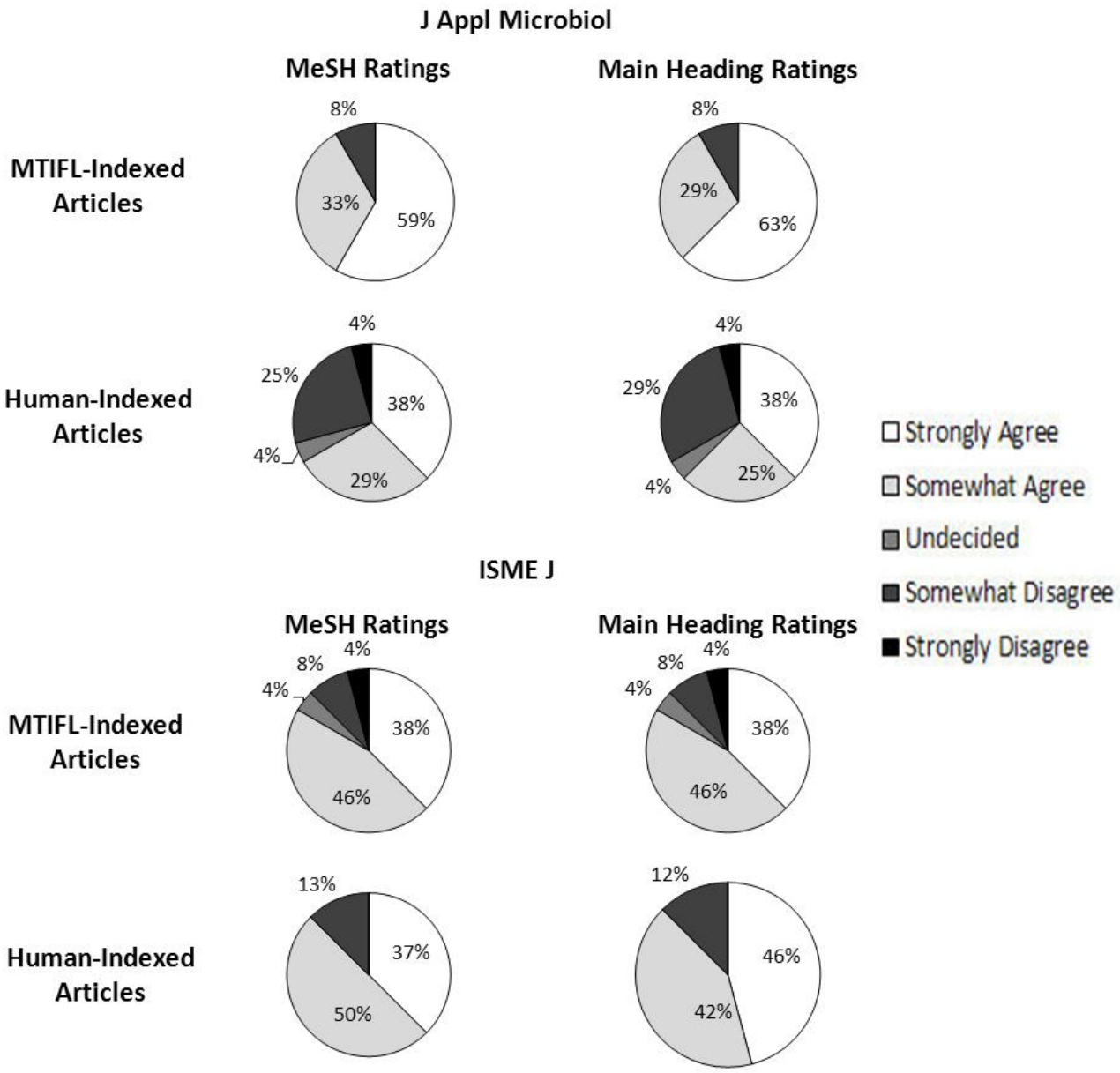

Figure 10. MeSH (charts on left) and IM (charts on right) term rating data for *J Appl Microbiol* (top panel) and *ISME J* (lower panel) for MTIFL-indexed (top portion of each journal panel) and human-indexed (bottom portion of each journal panel) articles.

Discussion

We tested three different approaches to determine a method that could be suggested for the evaluation of MTIFL indexing. While the sample sizes used in these tests were too small to complete an evaluation of MTIFL, the sample sizes were appropriate for an assessment of the approach. That is, generalizations about the conclusions that could be generated from these methods can be made based on these initial assessments. Overall, none of the three approaches tested were found to be appropriate because each approach had positive and negative aspects.

Evaluation of the Lister Hill Center Statistics Method

The statistics generated by LHC can be used to monitor MTIFL-indexed MeSH and IM terms and can be used to provide continuous feedback to the MTI program, creating a better tool that can be used for indexing assistance. These statistics can be generated fairly quickly and are relatively easy to understand and interpret. However, the statistics do not give information about the quality of the final indexing. The quantitative results are based on the human indexer making changes during the indexing process, which may or may not reflect good indexing outcomes. However, as a method for MTIFL evaluation, this quantitative method supplies high quality quantitative data such as counts and explanations for MeSH terms that are being added and/or deleted from the MTIFL indexing and has potential if used in combination with other methods of evaluation.

Evaluation of the Senior Indexer Analysis Method

Similar to the sample size utilized in the LHC statistical analyses, the sample size (n=6 articles for each of the two journals, N=12 articles total) was on the smaller side. However, we were able to see statistical significance in a number of analyses, so generalizations about the usability of this method for the evaluation of the MTIFL indexing could be made. First, the biggest disadvantage for this method was the labor-intensive procedure. The method utilizes the expertise of one or more senior indexers (i.e., for consistent data), who are well versed with the policies and rules of NLM indexing and who are knowledgeable about the biomedical field being indexed. Because of this, taking senior indexer time and effort away from regular duties to analyze completed indexed articles may not always be a practical or feasible option. In addition to the senior indexers doing the critical analysis of the MeSH terms, another person versed in statistical methods and evaluation is needed to compile, analyze and interpret the data. For this pilot study, two Senior Indexers took the time to analyze 12 articles, and the Associate Fellow analyzed and interpreted the data. Thus, this method may not always be a feasible option to which the Index Section can commit time, effort and money.

Despite all of the constraints with this method, some of the evaluative data resulting from the critical analysis was beneficial. This approach provided excellent in-depth analysis for MTIFL indexing (i.e., details about applying MeSH terms throughout the entire indexing process). For example, even though the number of MeSH terms added/deleted compared to the final number of MeSH terms indexed for a particular article was not always statistically significant, the number

of critical MeSH terms (as deemed by the in-depth Senior Indexer Evaluation approach) was found to be significant for the indexing quality of the article. Additionally, the details from the categories about why MeSH terms were added or deleted (e.g., *Too Tangential*, *3rd Tier Term*, *Too Specific/General*) can lead to better filters and triggers for LHC staff to incorporate into the software for better automated assistance. Therefore, this approach, although labor-intensive, was found to be extremely beneficial to the evaluation of the MTIFL indexing.

Evaluation of the Indexer Rating Method

The method of rating indexed MeSH and IM terms, as conducted by NLM indexers, holds some promise as a method for evaluating indexing. The survey rating model takes minimal indexers' effort (although, indexed citations will need to be rated by multiple indexers so reliable rating data can be generated). This method might be good for quick and continuous indexing evaluation. The evaluative results are not rich with details about the MeSH and IM terms indexed, but the method could supply information on whether an article is indexed well enough, or needs further attention. Any method of indexing (traditional or MTIFL) could be evaluated via this rating method. Mostly, this approach has promise for expansion with automated methods for sampling, collecting, and analysis. If this method were combined with other methods of evaluation, potentially powerful analyses for the evaluation of indexing could be completed.

Conclusions

After completion of the testing and assessing of these quantitative and qualitative approaches, it is now clear what the ideal model for the evaluation of MTIFL should:

- Supply both quantitative and qualitative data.
- Lend itself to straightforward data collection.
- Permit easy analysis.
- Be feasible.

The workload of the indexers in the Index Section is already extensive; an evaluation model for MTIFL must be practical (i.e., integrated with the indexers workflow) if it is to be justifiable (i.e., efforts, costs and time) and truly useful.

One option that could be pursued for indexing evaluation is a combined approach of two or more of the tested methods. As indicated, the statistics from the LHC provided quick and efficient information about MTIFL indexing, including details about the MeSH terms that were ultimately deleted or added by the indexers. These data provided not only quick information about the indexing process of the MTIFL journals, but also about the level of confidence by continuously checking and evaluating the F-scores. Combining the LHC statistics with the Indexer Ratings

could give a more complete picture of the indexing process. This type of continuous, random rating could lead to a system of flagging journals that may need special attention for review.

In pursuit of the goal for access to biomedical information, the National Library of Medicine has been at the forefront of the collection, organization and dissemination of the world's biomedical literature. In order to maintain its' high-level of integrity, new approaches to the evaluation of MTIFL indexing need to be considered. The pilot studies reported here have proved to be a good starting point into the assessment of possible evaluation models that can ensure continued excellence in MEDLINE indexing.

Acknowledgements

I would like to thank my project sponsors, Rebecca Stanger and Olga Printseva, for submitting this project and for allowing me to take liberties to expand beyond the original proposed project idea. I want to thank them for their assistance and advice throughout the project. I send many thanks to Jim Mork for statistics and assistance with data interpretation and to Carol Fisher, Mary Hantzes, Eduardo Tello and Janice Ward who assisted in the rating of MeSH terms. And a great big thank you to Lou Knecht and Rebecca Stanger for taking the time to edit and help revise the project presentation and report.

I am also grateful to Dr. Donald A.B. Lindberg, Betsy Humphreys and the National Library of Medicine for continued support of the Associate Fellowship Program. And I would like to thank Sheldon Kotzin, Joyce Backus for support throughout the fellowship year.

Finally, I would like to thank Lou Knecht and Torri Kellough for their mentorship and Kathel Dunn for her guidance and support. To my fellow 2011-2012 Associate Fellows, Bethany Harris, Jessi Van Der Volgen and Michele Mason-Coles, for comradeship throughout the Fellowship year. And last but not least, a big thank you to my husband, Ryan Spaulding.

References

1. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Stud Health Technol Inform. 2004;107(Pt 1):268-72.

2. Aronson AR, Mork JG, Lang F-M, Rogers WJ, Jimeno-Yepes AJ, Sticco JC. The NLM Indexing Initiative: Current status and role in improving access to biomedical information: Report to the Board of Scientific Counselors, National Library of Medicine, Bethesda, MD, April 5, 2012.

3. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.

4. Aronson AR, Mork JG, Lang FM, Rogers WJ, Neveol A. NLM Medical Text Indexer: A tool for automatic and assisted indexing. US National Library of Medicine, LHNCBC, Communications LHNCfB; 2008. Report No.: LHNCBC-TR-2008-002.

5. Aronson AR, Bodenreider O, Demner-Fushman D, Fung KW, Lee VK, Mork JG, Neveol A, Peters L, Rogers WJ. From indexing the biomedical literature to coding clinical text: Experience with MTI and machine learning approaches. Proceedings of the ACL '2007 Workshop "BioNLP". 2007:105-12.

Appendix A.

Raw data for statistics used in journal analyses.

Indexing	#Articles	#MeSH	#IM	#SCR	Avg MeSH	MeSH Stdev	Avg IM	IM Stdev
Arch Microbiol								
Indexer (Issue 1)	7	85	32	30	12.14	5.27	4.57	0.79
Indexer (Issue 2)	5	63	22	30	12.60	4.88	4.40	1.14
Indexer (Issue 3)	7	75	28	34	10.71	4.07	4.00	1.00
MTIFL Method (Issue 1)	7	84	26	34	12.00	1.91	3.71	1.11
MTIFL Method (Issue 2)	7	60	25	26	8.57	3.51	3.57	1.51
MTIFL Method (Issue 3)	8	98	34	44	12.25	3.11	4.25	0.71
Can J Microbiol								
Indexer (Issue 1)	14	200	66	63	14.29	2.89	4.71	1.54
Indexer (Issue 2)	16	153	70	48	9.56	3.24	4.38	1.20
Indexer (Issue 3)	9	108	34	26	12.00	3.67	3.78	0.67
MTIFL Method (Issue 1)	13	149	48	74	11.46	3.60	3.69	0.75
MTIFL Method (Issue 2)	10	133	43	27	9.57	3.20	4.57	0.67
MTIFL Method (Issue 3)	7	67	32	12	11.92	2.51	3.85	1.27
ISME J								
Indexer (Issue 1)	10	132	42	41	13.20	5.37	4.20	1.14
Indexer (Issue 2)	10	129	48	28	12.90	4.09	4.80	1.69
Indexer (Issue 3)	9	77	34	15	8.56	3.00	3.78	0.83
MTIFL Method (Issue 1)	16	146	68	38	9.13	2.47	4.25	1.61
MTIFL Method (Issue 2)	14	131	53	38	9.36	1.45	3.79	1.19
MTIFL Method (Issue 3)	12	112	42	31	9.33	2.10	3.50	1.17
J Appl Microbiol								
Indexer (Issue 1)	36	441	164	127	12.25	3.06	4.56	1.34
Indexer (Issue 2)	37	485	173	146	13.11	3.30	4.68	1.47
Indexer (Issue 3)	37	462	152	143	12.49	3.32	4.11	1.20
MTIFL Method (Issue 1)	27	344	115	101	12.74	2.96	4.26	1.02
MTIFL Method (Issue 2)	24	283	104	109	11.79	2.40	4.33	0.87
MTIFL Method (Issue 3)	25	282	110	82	11.28	3.14	4.40	1.00

Appendix B.

A description of MTI statistics.

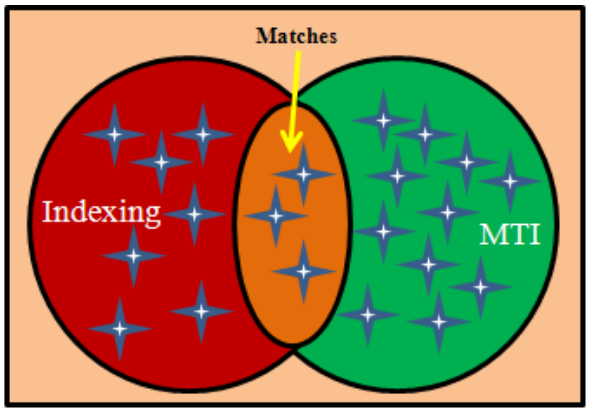

Figure 1 - Example for detailing Recall, Precision, and F2 calculations

The following table illustrates how to compute the Recall, Precision, and F_2 measure the way we do for the MTI statistics using the information in *Figure 1*.

Number of Indexing MHs	10
Number of MTI Recommendations	14
Number of Matches	3
Recall	$\textit{Number Matched} / \textit{Number Indexing MHs} = {}^{3}/_{10} = 0.3$ *(30.00%)*
Precision	$\textit{Number Matched} / \textit{Number MTI Recommendations} = {}^{3}/_{14} = .2143$ *(21.43%)*
F_2 Measure	$\frac{5 * Precision * Recall}{(4 * Precision) + Recall} = \frac{5 * .2143 * .3}{(4 * .2143) + .3} = .2778$ *(27.78%)*