# Evaluation of a simple method for the automatic assignment of MeSH descriptors to health resources in a French online catalogue

**Aurélie Névéol[a], Suzanne Pereira[b, c, d], Gaetan Kerdelhué[b], Badisse Dahamna[b], Michel Joubert [d], Stéfan J. Darmoni[b, c]**

[a] *U.S. National Library of Medicine, National Institutes of Health, Bethesda, USA*
[b] *CISMeF, Research Department, Rouen University Hospital, France*
[c] *GCSIS, LITIS EA 4051, Institute of Biomedical Research, University of Rouen, France*
[d] *LERTIM, Marseille Medical University, France*

**Abstract**

***Background****: The growing number of resources to be indexed in the catalogue of online health resources in French (CISMeF) calls for curating strategies involving automatic indexing tools while maintaining the catalogue's high indexing quality standards.* ***Objective****: To develop a simple automatic tool that retrieves MeSH descriptors from documents titles.* ***Methods:*** *In parallel to research on advanced indexing methods, a bag-of-words tool was developed for timely inclusion in CISMeF's maintenance system. An evaluation was carried out on a corpus of 99 documents. The indexing sets retrieved by the automatic tool were compared to manual indexing based on the title and on the full text of resources.* ***Results****: 58% of the major main headings were retrieved by the bag-of-words algorithm and the precision on main heading retrieval was 69%.* ***Conclusion****: Bag-of-words indexing has effectively been used on selected resources to be included in CISMeF since August 2006. Meanwhile, on going work aims at improving the current version of the tool.*

***Keywords :***

Abstracting and Indexing/methods; Algorithms, Catalogs, Library; Information Storage and Retrieval/methods; Evaluation Study, France; Medical Subject Headings; Natural Language Processing

## Introduction

Since 1995, the catalogue of online health resources in French (CISMeF)[1] has been selecting institutional and educational resources for patients, students and health professionals. The resources are described with a set of metadata and Medical Subject Headings (MeSH® descriptors[1]). Faced with a growing amount of resources to be indexed and included in the catalogue, the CISMeF team has recently adopted a new curation policy. The use of automatic indexing tools was deemed necessary to reduce the indexing back-log of about 7,000 resources. Considering the limitations of automatic indexing methods, it is necessary to distinguish clearly which resources in the catalogue are indexed manually and automatically [3]. In fact, the question arises at the time resources are considered for inclusion in the catalogue: the curation policy must define which resources should be indexed with the higher quality indexing produced manually and which resources may be given less attention and indexed automatically. At CISMeF, the decision was mainly based on two criteria: (a) the depth of indexing required and (b) the level of coverage of a given topic. High quality (manual) indexing must be available for all the topics covered in the catalogue, but when a reasonable level of coverage has been reached (about a dozen resources according to the curator) additional resources may be indexed automatically. Besides, we assume that automatic indexing is more suitable to cover only the central main concepts discussed in a resource whereas manual indexing will be necessary for in-depth indexing. In any case, in answer to an information query, manually indexed resources will always be displayed before automatically indexed resources. Moreover, the type of indexing (manual *vs.* automatic) will be shown to the user.

Teaching material (N=3629) and clinical guidelines (N=2978) are two vast categories of resources indexed in CISMeF. The average number of descriptors used to index a teaching resource is 9.89 +/- 10.87 vs. 13.64 +/- 15.69 for a clinical guideline. The difference is significant according to a Student test ($p<0.0001$) and illustrates the fact that clinical guidelines are indexed more in depth than teaching resources. In fact, for clinical guidelines, the curation policy for CISMeF is to be as exhaustive and as minute as possible: all clinical guidelines are to be indexed manually regardless of the current coverage in the catalogue. Teaching resources, on the other hand, may be given less attention once sufficient coverage has been reached, as they do not require in-depth indexing. To accommodate the difference between these resource types, a longer time is typi-

[1] MeSH is the US. National Library of Medicine (NLM)'s controlled vocabulary thesaurus. It consists of sets of terms naming biomedical related descriptors in a hierarchical structure. MeSH is used worldwide for indexing articles from the biomedical literature.

cally spent for indexing clinical guidelines2 (about 60 minutes *vs.* 15 minutes for teaching resources). Furthermore, CISMeF's curation policy is to use automatic indexing on teaching resources[3].

The VUMeF[4] project aims at increasing the amount of material available for French in the Unified Medical Language System. In particular, VUMeF participants agreed to make the development of automatic indexing tools a priority task of the project. It was addressed in two steps: first, automatically extracting a set of MeSH descriptors from online resources in French (which is discussed below) and secondly, in this set, selecting the major descriptors denoting central concepts discussed in the resource [2]. In the framework of VUMeF, the CISMeF team has been consistently researching and evaluating advanced automatic MeSH indexing techniques [3]. Although promising results have been obtained from a research prototype (see [4] for a comparative evaluation of CISMeF's MAIF and the NLM's Medical Text Indexer [5]), the integration of an operational system into the catalogue workflow is a lengthy process. For this reason, more readily available techniques are also investigated in order to speed-up the availability of automatic indexing tools and eventually complement other tools when they become available. Recent advances in Information Retrieval in CISMeF have resulted in the development of a query analysis algorithm designed to map free text to MeSH [7]. Moreover, new partnerships with health information providers accommodate the reception of resources to be included in the catalogue along with some metadata information such as the title, which previous research has shown to be significantly informative of the document content [3], [6]. In this paper we (a) assess the extent of the information contained in resource titles and (b) evaluate an automatic MeSH indexing approach based on bag-of-words indexing of resource titles in the specific context of teaching resources.

## Materials and Methods

### Bag-of-words indexing algorithm

The algorithm used to extract keywords from the title of documents is similar to the query interpretation algorithm currently used in CISMeF [7]. We decided to use this algorithm for indexing based on the assumption that document titles for teaching resources are similar to information retrieval queries with respect to length and information content. In this experiment, teaching resources titles are processed in the same way as queries to extract MeSH descriptors which will then be used as candidate terms to index the teaching resource.

After the title has been normalized (accents are removed, all words are switched to lower case…) and stop words have been removed, a bag of words containing all the content words is formed. Each word is stemmed in order to account for some cases of word flexion and derivation. The "bag" thus obtained is sorted alphabetically and matched against a database of MeSH terms that have been processed in the same way. If no term is found, subsets of the original bag of words are processed. The size of the bags is decreased by one at each step of the process. In an effort to retrieve the most specific keywords, when a match is found, the corresponding content words are taken out of the bag before the next iteration. For a given bag size, when more than one match is found all candidates are kept. For example, the title "prevalence of hepatitis A and B" would yield both *Hepatitis A* and *Hepatitis B* when size-2 bags are considered, but not *Hepatitis* which would have been removed before size-1 bags were processed. If both MeSH main headings and subheadings are retrieved, all the legal[5] pairs are formed from both main heading and subheading sets.

Figure 1 illustrates the processing of a sample title[6] in the test corpus (see next section for corpus description). A bag of seven content words is obtained from the title "Tumeurs cérébrales chez l'enfant: particularités épidémiologiques, diagnostiques et thérapeutiques" (*Brains tumors in children : epidemiologic, diagnostic and therapeutic specificities*). No single MeSH term can be matched to it. Therefore, the size of the bag is gradually reduced. The two-word bag containing "cerebrales" and "tumeurs" yields the main heading "tumeurs du cerveau" (*Brain Neoplasms*). The two corresponding content words ("cerebrales" and "tumeurs") are taken out of the bag. Size-one bags containing the remaining words yield the main heading "enfant" (*Child*) and the subheadings "épidémiologie" (*epidemiology*) and "thérapie" (*therapy*). Finally, as both subheadings are allowable qualifiers for *Brain Neoplasm* but not for *Child*, a list of three indexing candidates is produced: *Child*, *Brain Neoplasms/epidemiology* and *Brain Neoplasms/therapy*. This particular example will be further commented on in the discussion section.

### Test Corpus

The algorithm was evaluated on a corpus of 99 teaching resources to be included in CISMeF that were selected for automatic processing. A professional indexer was asked to provide MeSH indexing sets for the corpus resources. First, the indexer was only shown the title of the resource and the indexing set automatically produced using the bag-of-words algorithm. His task was to revise the indexing set obtained automatically. Then, the indexer had access to the full text of the resource and was asked to index the resource as he would usually do, i.e. he selected MeSH descriptors to index the resource, and assigned to each a "major" or "minor" weight depending on how substantively the concept represented by the descriptor

[2] Other cataloguing institutions may adopt different policies regarding time issues – e.g. at NLM, the average time spent indexing an article for Medline is 15 minutes regardless of the publication type.

[3] e.g. a resource discussing drugs available for the treatment of thrombosis was selected for automatic processing because it consisted of lecture notes and the query "thrombosis/drug therapy" retrieved 31 manually indexed resources on 11/08/06

[4] *Vocabulaire Unifié Médical Français* (French Unified Medical Vocabulary). Project sponsored by the French National Research Agency. http://www.vidal.fr/vumef/

[5] For each main heading, MeSH defines a set of subheadings called "applicable qualifiers" that can be coordinated with it (e.g. */metabolisms* is applicable to *Amino Acids* but not *Hand*).

[6] The original title ("Cancer de l'enfant: particularités épidémiologiques diagnostiques et thérapeutiques") was edited to illustrate an additional feature of the algorithm.

was discussed in the resource. As a result, each resource in the test corpus was manually annotated with two different sets of indexing terms: one based on the title only, and one based on the full text.

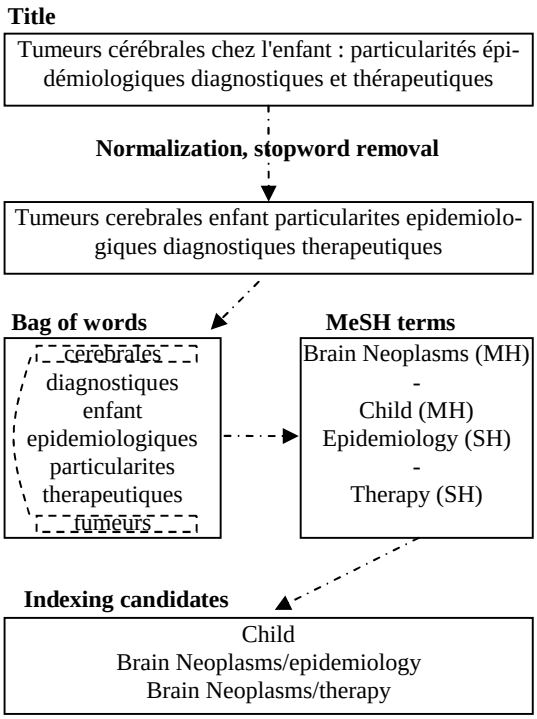


*Figure 1- Bag-of-words indexing of a sample title*

It is important to stress that the annotations were obtained through an iterative process in which the quality of the indexing improved with every step, while being consistent [8] with the previous step: the automatic bag-of-words recommendations were revised to obtain manual title annotations. In turn, these were revised to obtain manual full text annotations.

In this evaluation, the indexer is not blind to the recommendations produced by the automatic tool. However our goal was to conduct an experiment reproducing real-life indexing settings where the recommendations of the automatic tool will be available to the human indexer when indexing a resource.

The revisions made by the indexer include adding a main heading, deleting a main heading, modifying the list of subheadings attached to a main heading. Table 1 shows the successive steps for a sample resource in the test corpus.

### Evaluation measures

In this study we used precision and recall to compare a set of candidate indexing terms to a set of reference indexing terms. *Precision* corresponds to the number of terms present in both the candidate and reference sets over the total number of terms in the candidate set. *Recall* corresponds to the number of terms present in both the candidate and reference sets over the total number of terms in the reference set.

*Table 1 – Processing of a sample corpus resource: overview of the indexing revision cycle*

| | |
|---|---|
| **Title** | Cancer de l'enfant: particularités épidémiologiques diagnostiques et thérapeutiques<br>(*Cancer in children : epidemiologic, diagnostic and therapeutic specificities*) |
| **Bag-of-words indexing (Title)** | Neoplasms/epidemiology<br>Neoplasms/therapy<br>Child |
| **Manual indexing (Title)** | Neoplasms/diagnosis<br>Neoplasms/epidemiology<br>Neoplasms/therapy<br>Child<br>Pediatrics |
| **Manual indexing (Full Text)** | *Neoplasms/diagnosis<br>*Neoplasms/epidemiology<br>*Neoplasms/therapy<br>Child<br>Continuity of Patient Care<br>Pediatrics/education |

In addition, we considered three categories of terms:

- Descriptors: MeSH main headings or main heading /subheading pairs. In this category, subheading coordination is taken into account (e.g. *Pediatrics* does not match *Pediatrics/education*)
- Main headings: Any MeSH main heading. In this category, subheadings or stars (indicating major concepts) are not taken into account. (e.g. *Pediatrics* matches **Pediatrics/education*)
- Central-concept main headings: MeSH main headings that were marked as "major" using the star symbol "*" in the manual indexing based on the resource full text. In this category, subheadings are not taken into account. (e.g. **Pediatrics* matches **Pediatrics/education*)

## Results

### Information content of resource titles

*Table 2 – Information content of resource titles*

| | Information Content Precision (%) – Recall (%) | |
|---|---|---|
| **Descriptors** | 71 | 24 |
| **Main Headings (MH)** | 81 | 37 |
| **Central-concept (*MH)** | 78 | 78[7] |

The extent of the information contained in teaching resource titles was assessed by comparing the manual indexing obtained from the title of the resource to the manual indexing obtained

[7] Precision and recall figures are the same for central concepts because the indication of whether a term is central (major) is only available for manual indexing on the full text of the resource.

from the full text of the resource (see lines 3 and 4 in Table 1 for a specific example – in this case, the precision for *descriptors* indexing was 4/5=80% whereas the recall was 4/6=67%).

**Performance of the bag-of-words indexing**

The performance of the bag-of-words indexing was assessed by comparing the automatic indexing recommendations obtained by applying the bag-of-words algorithm on the resource title to the manual indexing:

- obtained from the resource title (overall results are shown in table 3; see lines 2 and 3 in table 1 for a specific example. In this case, the precision for *descriptors* indexing was 3/3=100% whereas the recall was 3/5=60%).

*Table 3 – Performance of bag-of-words indexing (Title)*

| | Performance Precision (%) – Recall (%) | |
|---|---|---|
| **Descriptors** | 62 | 56 |
| **Main Headings (MH)** | 69 | 64 |
| **Central-concept (*MH)** | 58 | 58 |

- obtained from the full text of the resource (overall results are shown in table 4; see lines 2 and 4 in table 1 for a specific example – in this case, the precision for *descriptors* indexing was 3/3=100% whereas the recall was 3/6=50%).

*Table 4 – Performance of bag-of-words indexing (Full Text)*

| | Performance Precision (%) – Recall (%) | |
|---|---|---|
| **Descriptors** | 54 | 16 |
| **Main Headings (MH)** | 66 | 30 |
| **Central-concept (*MH)** | 58 | 58 |

## Discussion

**Information content of teaching resources titles**

According to table 2, overall, only 24% of the MeSH descriptors to index a teaching resource may be inferred from the resource title, however including 78% of the central concepts. This shows that, in our corpus, teaching resources titles contain explicit information as to the central content of the resource. In 81% of the cases, the main headings inferred from the title by the indexer were kept after reviewing the full text of the resources. For descriptors, this figure goes down to 71%. Some descriptors were discarded for not denoting concepts that were substantively discussed in the resource. The other descriptors were in fact main headings to which a subheading had to be attached – in the example presented in table 1, the subheading *education* had to be attached to *Pediatrics*, which was inferred from the title.

**Bag-of-words indexing**

***Performance***

According to table 4, more than half (58%) of the major main headings were retrieved by the bag-of-words algorithm and the precision on main heading retrieval was 69%. These results show that the algorithm is able to retrieve central concepts, while generating a reasonably low noise. However, the difference between the performance on descriptors and main headings (lines 2 vs. 4 in tables 3 and 4) indicates that the more difficult task of assigning subheadings to the main headings is lacking. The low recall for descriptors and main headings compared to the full text (table 4) reflects the amount of information that may be extracted from the sole title of the resource. For example, for main heading retrieval recall could not be higher than 37% (as shown in table 2), so 30% is comparatively a good performance.

***Error Analysis***

Looking at sample revisions of the bag-of-words recommendations made by the indexer helps identifying the issues that need to be addressed in order to improve the automatic tool. Typical errors fall into the following categories:

- Stemming errors – in the example shown in table 1, the word « diagnostiques » was not stemmed properly and could not be mapped to the subheading *Diagnostic* (*Diagnosis*). As a result, the pair *Neoplasm/diagnosis* was not retrieved
- Generic main headings – some MeSH terms that may appear in a resource title are so generic that they are rarely used for indexing. *Syndrom*, *Patient*, *Life* or *Health* are examples of such descriptors.
- Implicit indexing rules – through daily indexing practice, the indexers are able to infer descriptors that do not explicitly appear in the title of the resource. In the example shown in table 1, no cue from the title prompts the use of the main heading *Pediatrics*. However, *Pediatrics* is typically an appropriate descriptor to index a teaching resource discussing a particular disease onset in children (here, cancer).
- Level of specificity – some of the descriptors retrieved by the algorithm were sometimes related to descriptors selected by the indexers, although not identical. For example, *Hemorrhage/therapy* was retrieved instead of *Gastrointestinal Hemorrhage/therapy*.

Based on these observations, a « stop list » of common generic descriptors is currently being compiled and shall be used to reduce the noise of the algorithm. As such, the bag-of-words algorithm cannot be expected to retrieve descriptors that an indexer would *infer* rather than see in a resource title. For this reason, applying post-processing rules (such as described in [9]) to a set of main headings retrieved by the algorithm would be desirable in order to improve the automatic tool.

Although we anticipated that the loss of word order inherent to the “bag-of-word” approach may yield some noise, no errors related to word-order were observed on the test corpus. This

may be explained by the fact that teaching resources titles are generally short and to-the-point. Different results may be obtained when applying the algorithm to more lengthy and complex sentences as can be found in full text.

**Limitations of this study**

The bag-of-words indexing algorithm presented here was evaluated on a set of 99 teaching resources. The small size of the evaluation corpus is due to the amount of manual labour needed to produce the annotations of the title and full text of the resources. Moreover, the fact that the indexer was shown the automatic recommendations while he was producing his own set of descriptors may yield a bias in favour of the automatic tool.

**Impact on CISMeF indexing procedure**

The positive results of this study conducted in May 2006 prompted the effective use of bag-of-words indexing in the CISMeF catalogue as of August 2006. The original back-log of 6,832 resources was automatically processed with the bag-of-words algorithm. The automatic indexing for 1127 resources (including the 99 resources of our corpus) have been revised by an indexer and included with the manually indexed resources of the catalogue. For another 557 resources, manual revision is pending. Finally, the remaining 5148 resources are included with the automatically indexed resources of the catalogue (these resources had originally been classified as "low priority" and did not require in depth indexing).

**Perspectives**

***Automatic indexing for CISMeF***: In the near future, the automatic indexing of selected low-priority resources to be included in the CISMeF catalogue will no doubt consist of integrating the recommendations produced by the different methods studied by the CISMeF Team: MAIF (i.e. a combination of Natural Language Processing and Nearest Neighbors approaches) and the bag-of-words indexing presented here.

***Other uses of bag-of-words indexing***: Based on the results of this study, we are planning to adapt the bag-of-words algorithm to the coding of patient discharge summaries with ICD-10 [10], MeSH and SNOMED [11] and to the indexing of French FDA notices with the Unified Vidal Thesaurus (TUV). In this case, documents would be segmented at the sentence level and the algorithm would be applied to each sentence. Ultimately, our goal is to integrate these applications to produce a multi-terminology indexing tool able to process medical documents and extract concepts belonging to several terminologies (MeSH, SNOMED, ICD-10 and TUV).

## Conclusion

We have presented a simple bag-of-words indexing method that retrieves MeSH descriptors from resource titles. An evaluation on a corpus of teaching resources shows good performance, in particular for central concepts. For this reason, bag-of-words indexing has been in use for selected CISMeF resources since August 2006.

## Acknowledgments

This research was supported in part by VIDAL (http://www.vidal.fr/) and by an appointment of A. Névéol to the NLM Research Participation Program sponsored by NLM and administered by the Oak Ridge Institute for Science and Education. The authors would like to thank CISMeF indexers for their help in the study design and result analysis.

## References

[1] Darmoni SJ, Leroy JP, Thirion B, Baudic F, Douyère M and Piot J. CISMeF: a structured Health resource guide. *Meth Inf Med* 2000, 39(1): 30-5.

[2] Joubert M, Peretti AL, Darmoni SJ, Dahamna B, Fieschi M: Contribution to an Automated Indexing of French-language Health Web Sites. *Proc. AMIA Symp 2005*, pp 409-13.

[3] Névéol A : Automatisation des tâches documentaires dans un catalogue de santé en ligne. PhD dissertation. INSA de Rouen. (2005).

[4] Névéol A, Mork JG, Aronson AR, Darmoni S.J : Evaluation of French and English MeSH Indexing Systems with a parallel corpus. *Proc. AMIA Symp. 2005*, pp 565-9.

[5] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *Proc. Medinfo. 2004*, pp. 268-72.

[6] Barthes R, Analyse textuelle d'un conte d'Edgar Poe, l'aventure sémiologique. Paris, Seuil, 1985.

[7] Soualmia LF : Etude et évaluation d'approches multiples d'expansion de requêtes pour une recherche d'information intelligente. PhD dissertation. INSA de Rouen (2004).

[8] Funk, M. E., Reid, C. A., & McGoogan, L. S. (1983). Indexing consistency in MEDLINE. *Bull. Med. Libr. Assoc.*, 2 (71):176-83.

[9] Névéol A., Shooshan SE., Humphrey SM., Rindflesh TC., Aronson AR. (2007): Multiple approaches to fine-grained indexing of the biomedical literature. *Proc. PSB.*

[10]World Health Organisation (1992). ICD-10 International Statistical Classification of Diseases and Related Health Problems: Tabular List v. 1. Material.

[11]Lussier YA., Rothwell DJ., Cote RA., The SNOMED model: a knowledge source for the controlled terminology of the computerized patient record. *Methods Inf Med.*, 1998 Jun;37(2):161-4.

**Address for correspondence**

Pr. Stéfan Darmoni - CHU de Rouen, DIR

1, rue de Germont - 76031 Rouen cedex - FRANCE