Evaluation of Meta-Concepts for Information Retrieval in a Quality-Controlled Health Gateway

Jean-François Gehanno [a], Benoit Thirion[a,b] , Stéfan J. Darmoni[a,b]

[a] GCSIS, LITIS EA 4108, Rouen University Hospital, France
[b] CISMeF, Rouen.University Hospital

Abstract

Background: CISMeF is a French quality-controlled health gateway that uses the MeSH thesaurus. We introduced two new concepts, metaterms (medical specialty which has semantic links with one or more MeSH terms, subheadings and resource types) and resource types.
Objective: evaluate precision and recall of metaterms.
Methods: We created 16 pairs of queries. Each pair concerned the same topic, but one used metaterms and one MeSH terms. To assess precision, each document retrieved by the query was classified as irrelevant, partly relevant or fully relevant.
Results: the 16 queries yielded 943 documents for metaterm queries and 139 for MeSH term queries. The recall of MeSH term queries was 0.44 (compared to 1 for metaterm queries) and the precision were identical for MeSH term and metaterm queries.
Conclusion: Metaconcept such as CISMeF metaterms allows a better recall with a similar precision that MeSH terms in a quality controlled health gateway.

Keywords

Abstracting and Indexing; Cataloguing; Controlled Vocabulary; France; Information Storage and Retrieval; Internet; Medical Subject Headings; Publication types; Subheadings.

Introduction

The Internet and in particular the Web has become an extensive health information repository. In this context, several quality-controlled health gateways have been developed [1]. Quality-controlled subject gateways were defined by Koch [2] as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery. Considerable manual effort is used to process a selection of resources which meet quality criteria and to display an extensive description and indexing of these resources with standards-based metadata. Regular checking and updating ensure optimal collection management. The main goal is to provide a high quality of subject access through indexing resources using controlled vocabularies and by offering a deep classification structure for advanced searching and browsing.

Among several quality-controlled health gateways, CISMeF ([French] acronym for Catalog and Index of French Language Health Resources on the Internet) [3] was designed to catalog and index the most important and quality-controlled sources of institutional health information in French in order to allow end-users to search them quickly and precisely (N=23,948). CISMeF (http://www.cismef.org) is manually indexed by a team of four indexers, which are medical librarians and systematically checked by the chief information scientist (the “super-indexer”).

CISMeF uses two standard tools for organizing information: the MeSH thesaurus [4] and several metadata element sets, in particular the Dublin Core metadata format (URL:http://www.dublincore.org) [5]. The use of specific metainformation is crucial in order to improve the recall and precision of internet searches [6]. As proposed by Hoelzer et coll. [6], CISMeF uses XML and RDF to meet these requirements. This structure enables us to place the project at an overlap between the actual informal Web and the forthcoming Semantic Web.

However, the MeSH thesaurus was originally intended to index scientific articles for the Index Medicus and for the MEDLINE database. In order to customize it to the broader field of health Internet resources, we have been developing several enhancements [3] to the MeSH thesaurus, with the introduction of two new concepts, metaterms (MT) and resource types (RT) respectively. The CISMeF terminology is shown in Figure 1. CISMeF resource types (RT) are an extension of the publication types of MEDLINE.

A metaterm is a medical specialty or a biological science (e.g. *cardiology, bacteriology)*, which has semantic links with one or more MeSH terms, subheadings and RTs. To construct a taxonomy of medicine, the publishing division of the American Medical Association (AMA) took as its precedent the simplified access to MeSH via CISMeF metaterms [6].

The goal of this article was to assess the precision and recall of metaterms compared to MeSH terms in CISMeF quality-controlled subject gateway.

Materials and Methods

CISMeF terminology

In 2007 the CISMeF team is composed of four medical librarians, two medical informaticians, one engineer, and three PhD students majoring in Computer Science.

The CISMeF terminology is exploited for several tasks: resource indexing performed manually, resource categorization performed automatically, visualization and navigation through the concept hierarchies and a CISMeF Terminology Server (URL: http://www.chu-rouen.fr/terminologiecismef/) and information retrieval using the Doc'CISMeF search engine.

The MeSH was selected because it responds to the aims of the medical librarians and it is well known by the health professionals. Approximately 24,357 MeSH terms (e.g.: *abdomen, hepatitis*) and 83 qualifiers (e.g.: *diagnosis, complications*) compose the MeSH thesaurus in its 2007 version. These concepts are organized into hierarchies going from the most general on at the top of the hierarchy to the most specific in the bottom of the hierarchy. For example, the MeSH term *hepatitis* is more general than the MeSH term *hepatitis viral A*. The qualifiers, also organized into hierarchies, allowing to specify which particular aspect of a keyword is addressed, and then to focus on a sub-field of the keyword. For example the association of the keyword *hepatitis* with the qualifier *diagnosis* (noted *hepatitis/diagnosis*) restrict the *hepatitis* to its *diagnosis* aspect. The "*is-a"* relations between concepts are extracted from the MeSH text files to define the subsumption relationships in the CISMeF keywords hierarchy.

A CISMeF metaterm is a medical specialty or a biological science (e.g. *cardiology, bacteriology)*. In fact, the idea of creating meta-terms came up to optimize information retrieval in CISMeF (Doc'CISMeF search engine; URL: http://doccismef.chu-rouen.fr/servlets/Simple) and to cope with the relatively restrictive nature of these medical specialties as MeSH terms. The MeSH thesaurus does not allow to have a global vision of a medical specialty. Therefore, in the CISMeF terminology, metaterms can be considered as "meta-concepts". Metaterms have been manually selected by the chief medical librarian (BT). The semantic links between metaterms and MeSH terms, MeSH subheadings and CISMeF resource types are based on his know-how and expertise of medical specialists of the Rouen University Hospital. There is a 0 to N relations between CISMeF metaterms and MeSH terms, MeSH subheadings and CISMeF resource types (see Figure 1). Each metaterm has a semantic link with the corresponding MeSH term, *e.g.* the metaterm *cardiology* has a semantic link with the MeSH term *cardiology*. For instance, the queries 'guidelines in cardiology' and 'databases in psychiatry' where *cardiology* and *psychiatry* are only MeSH keywords get few or no answers.

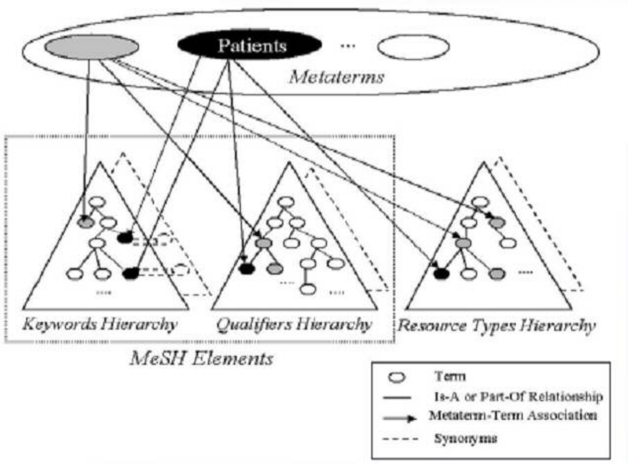

Figure 1: Semantic links between CISMeF metaterms and MeSH terms, MeSH subheadings and CISMeF resource types

Introducing *cardiology* and *psychiatry* as metaterms is an efficient strategy to get more results because instead of exploding one single MeSH tree (e.g. *psychiatry* as a MeSH term), using metaterms results in an automatic expansion of the queries by exploding other related MeSH, such as *psychiatric hospital* that belongs to a completely different tree structure within the MeSH, or CISMeF trees (for resource types) as well as the current tree (e.g. *psychiatric hospital* as a MeSH keyword or *mental health* dispensary as a resource type will be exploded in the case of the *psychiatry* query). In example, the metaterm *psychiatry* has the following semantic links: [MeSH Terms] "behavioral symptoms"; "community mental health centers"; "diagnostic and statistical manual of mental disorders"; "hospitals, psychiatric"; "mental disorders"; "mental health services"; "mentally ill persons"; "psychiatric department, hospital"; "psychiatric somatic therapies"; "psychiatric status rating scales"; "psychiatry"; "psychological techniques"; "psychophysiologic disorders"; "psychotherapy"; "psychotropic drugs" "schizophrenic psychology"; [CISMeF resource types] "community mental health centers"; "hospitals, psychiatric";

As defined by the Dublin Core Metadata Initiative (URL: http://www.dublincore.org/documents/dcmi-terms/) [5], a CISMeF resource type is used to categorize the nature or genre of the content of the resource. MeSH (term/subheading) pairs describe the topic of the resource. resource type is one of the fifteen Dublin Core repeatable and optional elements. For example, in the case of a clinical guideline about carbon monoxide intoxication, ‘carbon monoxide poisoning’ is the MeSH term and ‘clinical guidelines’ is the resource type.

In January 2007, the number of metaterms in the CISMeF terminology was 110. The comprehensive list of metaterms is available at the following URL: http://doccismef.chu-rouen.fr/liste_des_meta_termes_anglais.html.

Major Topics exist in the Medline database and the CISMeF catalogue for keywords and qualifiers. A term is said to be “major” if the concept it represents is discussed throughout the whole document, or on the contrary "minor" if it is referred to only in a few paragraphs. Major terms are marked in Medline & CISMeF by a star. In CISMeF, Major Topics are extended to resource types and metaterms. This task is manually performed by CISMeF medical librarians for resource types, and automatically performed for metaterms : a metaterm is “major” for a CISMeF resource if and only if at least one keyword, qualifier or resource type semantically linked to this metaterm is major for the same CISMeF resource (otherwise, the metaterm is minor).

In a comparative study performed by Abad Garcia *et al*. [1] among six European health gateways. Although CISMeF was rated second, it has been criticized because “failure on precision may be due to exhaustive indexing” [1]. To optimize the precision of our health gateway, we have introduced a major modification of the CISMeF information retrieval algorithm: when a query will be mapped to one or several terms of the CISMeF terminology (CISMeF metaterms, MeSH terms, MeSH subheadings, CISMeF resource types), the resources with Major Topics will be first displayed (e.g. in case of the following query 'guidelines in cardiology', resources with Major Topic cardiology as a metaterm and Major Topic guideline as a resource type will be first displayed).

Information retrieval queries

The objective of this article was to assess the precision and recall of metaterms compared to MeSH terms in a quality-controlled subject gateway. For this purpose, the CISMeF librarian team has created 16 pairs of queries. Each pair concerns the same topic, but with two different queries. One query contains at least one metaterm and one CISMeF resource type.

The other query contains at least one MeSH term and one CISMeF resource type (and no metaterm).

As an exemple, the query “guidelines in cardiology” is mapped in the CISMeF terminology as “guidelines[CISMeF resource type] AND cardiology[CISMeF metaterm]. This query will be compared to the following query: “guidelines[CISMeF resource type] AND cardiology[MeSH term].

Since each metaterm has a semantic link with the corresponding MeSH term, the query results for MeSH terms are all included in the query results for metaterms. Therefore, the golden standard for recall is provided by the query using the metaterm (Recall=1), to which query using MeSH terms are compared.

To assess precision, we examined each document proposed as a result of the query and classified it into three categories: irrelevant, partly relevant or fully relevant. We computed two Precision: one including all relevant (partly + fully) documents (P1), and one including only fully relevant documents (P2).

Therefore, we also calculated two Recall for MeSH term queries, one including all relevant (partly + fully) documents (R1) and an other one limited to fully relevant documents (R2).

The evaluation was performed by a physician from the LITIS Lab (JFG). To avoid bias, this physician does not belong to the CISMeF team. For each query, the physician evaluated the precision and the recall for the top 20 resources, because 95% of the end-users do not go beyond this limit when using search engines [8].

Results

Overall, the 16 queries yielded 943 documents for metaterm queries and 139 for MeSH term queries. Since we chose to assess relevancy of the first 20 documents, 304 documents were examined, 212 provided by the metaterm queries, and 92 by the MeSH term queries.

Detailed results are presented in table 1.

All the metaterm queries gave more results than MeSH term queries.

It is worthwhile to note that except for one query (teaching ressources for lung diseases), the absolute relevance was better with MT than with MeSH terms.

The recall for MeSH term queries, R1 and R2, were identical: 0.44.

The precision P1 was 0.60 and 0.61 for MT and MeSH term queries respectively.

The precision P2 was 0.46 and 0.47 for MT and MeSH term queries respectively.

Discussion

The goal of this article was to evaluate the precision and the recall of the metaterms compared to MeSH terms in information retrieval in a quality-controlled subject gateway.

PubMed has recently introduced a new tool to facilitate information retrieval by limiting the query to some specific subsets. These PubMed Subset Strategies are limited to 8 topics, which cover very different fields, from diseases (e.g. AIDS) to medical specialties (e.g. Complementary Medicine).

There are two main differences between CISMeF metaterms and PubMed Subset Strategies. First, PubMed Subset Strategies have a broader approach using MeSH terms and subheading, Journal names, Textword (which mainly includes Title and Abstract), which leads to a better recall. On the opposite, CISMeF metaterms have a better precision. Second, CISMeF metaterms offers a much better coverage than PubMed Subset Topics, since the latter includes only eight topics, as compared to the 110 topics included in CISMeF metaterms.

As expected, the recall of MeSH terms is low (R=0,44). Nevertheless, this does not imply a significant improvement of precisions, which were very similar between both queries (0.46 for MT vs 0.47 for MeSH term).

We only analyzed the first 20 results, which means that we examined 66% (92 out of 139) of documents proposed by MeSH terms queries, and 22% (212 out of 943) of documents proposed by MT queries. Although this could be considered unfair, we think this method reflects the real life, since most users usually don't go beyond the first web page or the first 20 documents [8].

CISMeF Metaterms already have two other use in information retrieval:

- First, in the CISMeF environment. They provide specific catalogues in different specialties: i.e. CISMeF team has provided a specific search engine in medical oncology for the French National Cancer Institute (URL: http://www.e-cancer.fr/) and another search engine in the disability domain (URL: http://doccismef.chu-rouen.fr/servlets/PIH). To design these two Web sites, we have used the following metaterms medical oncology and disability to limit any query (e.g. "asthma"[MeSH term] and "medical oncology"[Metaterm]).
- Second, in the MEDLINE bibliographic database via the PubMed Website (URL: http://www.pubmed.gov). Any end-user may use the list of metaterms available at the following URL: http://doccismef.chu-rouen.fr/liste_des_meta_termes_anglais.html to perform queries with a better recall as shown in the CISMeF catalogue (e.g. The query "wood toxicity in occupational medicine" is manually mapped to wood/toxicity[MeSH term] AND occupational medicine[Metaterm].

Several improvements of the semantic links between CISMeF metaterms and MeSH terms, MeSH subheadings and CISMeF resource types have been implemented: first with the help of several medical experts from the Rouen University Hospital; second and most important, with the help of the Network of National Library of Medicine (NNLM; URL: http://nnlm.gov/), using the Medlib-L listserv (URL: http://nnlm.gov/). Several medical librarians of the NNLM proposed some improvements to the MTs semantic links: they mostly helped to reduce the false negative results, when proposing new semantic links with MeSH terms and subheadings.

Conclusion

Queries using Meta-concept such as CISMeF metaterms increased recall for the first 20 items retrieved without sacrificing precision in a quality controlled health gateway.

References

[1]. Abad Garcia F, Gonzalez Teruel A, Bayo Calduch P, de Ramon Frias R, Castillo Blasco L. A comparative study of six European databases of medically oriented Web resources. J Med Libr Assoc. 2005;93(4):467-79.

[2]. Koch T. Quality-controlled subject gateways: definitions, typologies, empirical overview, Subject gateways. Online Information Review. 2000:24(1):24-34.

[3]. Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy JP et al.. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J. 2004:21(4):253-61.

[4]. Nelson SJ, Johnson WD, Humphreys BL. Relationships in Medical Subject Headings in Relationships in the organization of knowledge. In Bean and Green, eds. Kluwer Academic Publishers, 2001;171-84.

[5]. Dekkers M, Weibel S. State of the Dublin Core Metadata Initiative. D-Lib Magazine. 2003: 9(40).

[6]. Hoelzer S, Schweiger RK, Boettcher H, Rieger J, Dudeck J. Indexing of Internet resources in order to improve the provision of problem-relevant medical information. Stud Health Technol Inform. 2002;90:174-7.

[7]. Mc Gregor B. Constructing a concise medical taxonomy. J Med Libr Assoc. 2005; 93(1):121-3.

[8]. Spink A, Jansen BJ. Web search: Public searching of the web. 1^{rst} ed. Dordrecht: Kluwer Academic Publishers, 2004; 199p.

Table 1 : Comparison of the Metaterms or MeSH terms queries' results and relevancies

	N		Relevance of the first 20 documents*					
			0		1		2	
Query	MT	MH	MT	MH	MT	MH	MT	MH
A	1	0	0	0	0	0	1	0
B	4	0	0	0	3	0	1	0
C	45	23	9	9	1	2	10	9
D	38	1	2	0	3	1	15	0
E	11	0	0	0	2	0	9	0
F	256	36	3	1	1	2	16	17
G	15	2	11	0	3	1	1	1
H	36	0	11	0	8	0	1	0
I	90	0	7	0	1	0	12	0
J	10	8	1	0	0	0	9	7
K	21	14	10	7	7	6	3	1
L	375	48	17	17	1	1	2	2
M	4	4	0	0	0	0	4	4
N	5	3	2	1	0	0	3	2
O	2	0	0	0	0	0	2	0
P	30	0	12	0	0	0	8	0
Total	943	139	85	36	30	13	97	43

Queries :

A. Bibliographic databases in oncology

B. Toxic effects of occupational wood dust exposure

C. Recommendations for breast neoplasm therapeutics

D. Associations of patients suffering from heart diseases

E. Traumatology departments & units

F. Online lung diseases courses

G. Treatment of pain in children with cancer

H. Occupational neoplasm's' epidemiology in France

I. Patient information leaflets in digestive surgery

J. Libraries specialized in psychiatry

K. Guidelines for children's nursing pediatric units

L. Anatomical illustrations

M. Medical history museums in France

N. Periodicals in Urology

O. Nephrology clinical cases

P. Associations of patients suffering from addictions

N : Total number of documents retrieved

MT : Metaterms.

MH : MeSH terms.

* Relevance of retrieved documents : 0: irrelevant; 1: partly relevant; 2 : fully relevant

Address for correspondence

SJ. Darmoni, CISMeF, Rouen.University Hospital, 1 rue de Germont, 76031 Rouen Cedex, France