MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Doc’CISMeF: A Search Tool Based on “Encapsulated” MeSH Thesaurus

Stefan J. Darmoni[a c], Benoit Thirion[b], Jean Philippe Leroy[a], Magaly Douyère[a], Benoit Lacoste[c], Christophe Godard[c], Isabelle Rigolle[c], Martial Brisou[c], Stéphane Videau[c], Eric Goupy[a], Josette Piot[b], Myriam Quéré[a], Saïda Ouazir[b], Habib Abdulrab[c]

(a) Computer and networks department, Rouen University Hospital, France
(b) Medical library, Rouen University Hospital, France
(c) Perception and Information System Lab, National Institute of Applied Sciences, Mont Saint Aignan, France

Abstract

In the year 2000, the Internet became a major source of health information for the health professional and the Netizen. The objective of Doc’CISMeF (D'C) was to create a powerful generic search tool based on an structured information model which ‘encapsulates’ the MeSH thesaurus to index and retrieve quality health resources on the Internet. To index resources, D’C uses four sections in its information model: 'meta-term', keyword, subheading, and resource type. Two search options are available: simple and advanced. The simple search requires the end-user to input a single term or expression. If this term belongs to the D'C information structure model, it will be exploded. If not, a full-text search is performed. In the advanced search, complex searches are possible combining Boolean operators with meta-terms, keywords, subheadings and resource types. D’C uses two standard tools for organising information: the MeSH thesaurus and the Dublin Core metadata format. Resources included in D’C are described according to the following elements: title, author or creator, subject and keywords, description, publisher, date, resource type, format, identifier, and language.

Keywords:

Abstracting and indexing; Cataloging; Database; Information Storage and Retrieval; Internet; Subject Headings; Support, Non-U.S. Gov't; Vocabulary controlled.

Introduction

In the year 2000, the Internet became a major source of health information for the health professional and the Netizen [1]. For these two populations, access to accurate information on the Internet can be problematic; therefore, there is a great number of directories and search engines available in this new media [2]. However, directories, such as Yahoo [http://www.yahoo.com], or search engines, such as Altavista [http://www.altavista.com] do not permit the end-user to obtain a clear and organised range of available useful health information.

The objective of CISMeF [3] (Catalog and Index of French-speaking resources) was to assist the health professional during the search for electronic information available on the Internet. CISMeF was a project originally initiated by Rouen University Hospital (RUH). Its Universal Resource Locators (URL) are http://www.chu-rouen.fr/cismef or http://www.cismef.org. CISMeF began in February 1995 after the creation of the RUH’s Web site.

The major drawback of CISMeF was its limited technical capacity. It uses only static HyperText Markup Language (HTML) and does not use a database which allowed more complex searches, e.g. guidelines for hepatitis, combining the explode command not only for keywords and qualifiers but also for meta-terms and resource types. Therefore, in late 1999 the CISMeF team decided to create Doc'CISMeF (D’C).

The objective of D’C was to create a powerful generic search tool using the CISMeF information structure model which is based on an ‘encapsulated’ MeSH thesaurus. As with CISMeF, D’C was initiated by the Rouen University Hospital (RUH). Its URL is http://doccismef.chu-rouen.fr/. The first prototype version of D’C has been available on line since June 2000.

CISMeF covers healthcare disciplines and medical sciences. However, the scope of the D’C first prototype has originally a much narrower scope. It now covers practice guidelines, consensus conferences, technical reports and teaching materials as well as all the new resources included in CISMeF since August 4, 2000. When completed in September 2001, D’C and CISMeF will share the same scope and the same number of resources.

Material and methods

Hardware and software

The prototype version of D’C was implemented in June 2000 on a Wintel machine. D’C is a Web Site using an Apache http server (Apache server and Apache Jserv for the servlet’s execution). The D’C search tool is developed in Java Servlets because Java is an open and inexpensive solution. Furthermore, Java servlets are more efficient, easier to use, more powerful, more portable, and more cost effective than traditional CGI and many alternative CGI-like technologies.

The database is MS Access. The number of resources suggested this choice: approximately 10,000 at the end of the year 2000. If required (number of resources > 20 000), it will later be changed to Oracle. In this instance, because of Java portability, D'C can be transferred to a Unix platform without any major modifications. The Java servlets generate SQL requests via ODBC drivers, which read the Access database. Our database choices resulted from the general use of Oracle & Access in the RUH Hospital Information System. However, My SQL and PHP perform database applications more efficiently and are also an easier development environment.

The Webtrends Log Analyser, version 4.5.2. programs evaluated the use of the Web page after excluding it has excluded any requests made by computers registered to the RUH.

Standards and information model

D’C uses the same two standard tools as CISMeF for organising information: the MeSH (Medical Subject Headings) thesaurus from the US National Library of Medicine and the Dublin Core metadata format [4].

The MeSH thesaurus contains 19,771 MeSH terms and 83 qualifiers in its year 2000 version as well as nine classification levels. This thesaurus is accurate, rigorous and updated annually. We also use the French translation of this thesaurus, performed by the French Medlars Centre, the National Institute for Health and Medical Research (INSERM, and more specifically the DISC-DOC Network).

MeSH subheadings allow a focus on a sub-field of a MeSH term, e.g., chloride/toxicity. A French translation of the MeSH subheadings is also used. However, MeSH terms and qualifiers were less systematically used in CISMeF than in Medline.

Updating the D'C dynamic database is easier than the static CISMeF, therefore the resources included in D'C are more accurately described with additional MeSH terms (mean = 4.1 vs. 1.3 MeSH term per resource) and qualifiers than previously in CISMeF. In D’C, each keyword is no longer a 'de facto' MeSH Major Topic.

The information structure model of D’C ‘encapsulates’ the MeSH thesaurus with two semantic levels: meta-terms [5] (n=57) and resource types [6] (n=101).

The D’C and CISMeF 'meta-term' [5] is generally a medical specialty or a biological science, e.g., cardiology or bacteriology. These medical specialties are in most cases MeSH terms. The idea of the meta-term was established to cope with the relatively restrictive nature of these MeSH terms when searching 'guidelines in cardiology' or 'databases in virology' where cardiology and virology are meta-terms and guideline and databases are resource types.

In CISMeF, meta-terms had semantic links only with MeSH terms or categories. For example, on the 'oncology' page, the sites of general interest on this specialty were indexed and described, using the MeSH keyword 'medical oncology', followed by a list of starting points of related categories and other associated MeSH terms. The categories for oncology, were: (a) antineoplastic agents, (b) medical oncology, (c) neoplasms, and (d) tumors markers, biological. The MeSH term is: oncology service, hospital.

In designing D’C, we have added semantic links for meta-terms with the two other sections of our information model: qualifiers and resource types. The meta-term "oncology" has links with the subheading 'secondary' and the resource type 'oncology service, hospital'. The list of meta-terms is available at the following URL http://www.chu-rouen.fr/ssf/santspe.html.

D’C resource type is a expanded model of the publication type of Medline. We have added types which are specific to the resources available on the Internet, such as association, patient information and community networks. The list of resource types is available at the following URL: http://www.chu-rouen.fr/documed/typeressource.html.
Resource types describe the nature of a resource and MeSH terms describe the subject of a resource. For example, in the case of clinical guidelines regarding carbon monoxide intoxication, ‘carbon monoxide poisoning’ is the MeSH keyword and ‘clinical guidelines’ is the resource type.

Update of the Doc’CISMeF database

D'C shares the same method used in constructing the catalogue as CISMeF which involves a four-fold process: resource collection, filtering, description and index. One deputy medical librarian performs the resource collection and the information follow-up. The editorial board created in December 1996 [3] filters and selects the resources. Two deputy medical librarians describe and index the resources. The chief medical librarian is a ‘super-indexer’ in charge of validating the indexing. There is a daily 30-minute meeting with the medical informaticians for double-checking.

In order to include only reliable resources, D’C and CISMeF uses the main criteria (e.g. source, disclosure, last update) of Net Scoring [7] to assess the quality of health information on the Internet. There are 49 criteria which fall into eight categories: credibility, content, hyperlinks, design, interactivity, quantitative aspects, ethics and accessibility (see the list in English at the following URL: http://www.chu-rouen.fr/dsii/publi/netscoring.html). Some resources are not introduced in D’C and CISMeF because they do not respect basic, particularly ethical, criteria.

After the filtering process each resource is entered into the Access database and is described by using the 10 following elements from the 15 of Dublin core project [URL: http://purl.org/DC/about/element_set.htm]: title, author or creator, subject and keywords, description, publisher, date, resource type, format, identifier, and language. The following fields are specific to CISMeF: institution, city, province or state, country, target, cost of access, type of sponsorship.

Each resource in the database automatically generates: (a) the CISMeF record [3] previously hand-made; (b) the D'C complete record available in HTML and XML. This automatic generation uses stylesheets. The keyword section of the D'C complete record allows the end-user to return a search by clicking on a MeSH term or qualifier. In both cases, this search uses the explode command unlike Ovid [http://www.ovid.com] which is a commercial company that integrates bibliographic databases, such as Medline and full-text electronic journals. In the same keyword section, in addition to each MeSH term, a button is available to access the CISMeF MeSH page which contains the categories, the 'see also' relations and the synonyms in French. Each D'C complete record contains Dublin Core (DC) and CISMeF metadata [8].

Results: Description of Doc’CISMeF

D’C is an efficient and end-user-friendly search tool to find French-speaking worldwide health resources on the Internet. This Web site is principally and initially oriented for health professionals, although the general public may also have access to it. There is no restricted access in the D’C Web site. Using the Advanced Search, the end-user can select the documents designed for the general public ('target' field).

Whereas CISMeF, D’C has three priority axes: evidence based medicine, teaching material and patient information. D’C allows two types of searches in its prototype version: Simple Search and Advanced Search.

Simple Search

When using the 'Simple Search', the end-user may enter only one term or expression. This choice was driven by previous results during teaching courses with patients associations on how they used the CISMeF internal search engine and by the analysis of the CISMeF log regarding the internal search engine.

D’C allows bilingual searches in French (with or without accents) and in English, whether the end-user types in capitals or lower case letters (only for 'reserved' terms). A term is defined as 'reserved' if and only if it belongs to at least one section of the 'encapsulated' MeSH thesaurus: meta-term, MeSH terms, qualifiers or resource type. In this case, the 'reserved' term is exploded, meaning that all resources indexed in any of the situated terms below this term will be selected. If the 'reserved' term belongs to several levels of D’C information structure model (e.g. echography is a MeSH term and a qualifier), D'C unifies all the exploded levels as shown in the following equation (1):

$$\bigcup_{i=1}^{4} \exp(x) \qquad (1)$$

where i is the level of the information structure model, x is the 'reserved' term and exp is the explode function. For example, if the end-user enters the term 'echography' in English or 'échographie' in French, the search will be performed on the exploded Mesh term 'echography' and the exploded subheading 'echography'.

If the end-user is not typing a 'reserved' term, a full text search is performed on the most significant D’C fields: author, city, country, description or abstract, meta-term, keyword, publication date, qualifier or subheading, publisher, resource type, URL. We suggest that end-user employ two tools to help him to find the best MeSH terms according to his search: the MeSH Browser (URL: http://www.nlm.nih.gov/mesh/MBrowser.html) from the US NLM and its French translation by INSERM. The search allows left and right truncate by default.

If the search generates no results, a complementary search is proposed in the CISMeF internal search engine. If the term entered belongs to the MeSH thesaurus, a complementary search is proposed in PubMed.

Advanced Search

The Advanced Search can be performed on all main fields of the Doc’CISMeF database using boolean operators AND, OR and NOT

The explode command is activated by default. However, it is possible to deactivate this command if the end-user wants to focus on a MeSH term (or a qualifier or a resource type).

Common properties of Simple and Advanced Searches

For both searches, the results are displayed as a list of records. Four different displays are possible: (1) title, (2) title and URL, (3) title, URL, keywords, qualifiers and resource types and (4) title, URL, keywords, qualifiers, resource types and description, so called 'abridged record' (default choice). Whatever the choice of display, the access to the complete record is proposed for every record of the list. The number of records is also a variable parameter: 10, 20, 50 (default choice), 100. These resources are displayed either by date of publication in antichronological (default choice) as in PubMed, in alphabetic order using the title of the resources, by resource types or by country and city.

A complementary search in PubMed is proposed and automatically performed if the search includes terms from the MeSH thesaurus: MeSH terms, qualifiers and D'C resource types which are also Medline publication types, e.g. guidelines and technical reports. In the simple search, a meta-term will be translated into its equivalent MeSH term to perform the complementary PubMed request. In the advanced search, no PubMed search is performed if the end-user chooses a meta-term. This PubMed link is located at the end of each Doc’CISMeF result page.

A tutorial session describing the features of D’C is available at the following URL: http://doccismef.chu-rouen.fr/aide.html. According to the previous grant of CISMeF with AUF (Agence Universitaire de la Francophonie), the CISMeF team is responsible for the support of MDs and medical librarians in all French-speaking developing countries.

In some rare cases, one resource cannot be accurately indexed using the MeSH thesaurus: then, we use a 'manual

mapping' to search the nearest MeSH term, e.g. a resource coping with dysmelia has been indexed with the MeSH term ectromelia.
Currently the D'C prototype contains 5,700 resources whilst CISMeF contains over 10,100. We plan to complete the D'C database in September 2001. Then, D'C and CISMeF will contain the same number of resources (approximately 11,000). To index these resources, D'C uses only 4,100 MeSH terms (around 21% of the MeSH thesaurus).

Use patterns of the Web site

Our Web server software, which provides documents to users on request, does not know the identities of individual users, such as E-mail; the only identifying data available are the Internet IP addresses of the machines from which the users connect to the site.

During the month of September 2000, 4278 unique users made 18,418 requests. The respective percentage of Simple and Advanced searches were 87% (n=16,104) and 13% (n=2,314). These statistics underestimate the real figures due to the practice of file caching.

Discussion

D'C is a generic tool. It can be applied to English and to any other language where the MeSH has been translated (Finnish, German, Italian, Portuguese, Russian (transliterated), and Spanish [9]). The functions of D'C have been largely inspired by the PubMed and Ovid Web sites in the medical field and its graphic chart of simple and advanced search pages by the German site Geo-Guide (URL: http://www.geo-guide.de/) which is a geographical catalog.

In fact, the main strength of D'C relies on its information structure model. D'C conceptually encapsulates the MeSH structure (category, keyword, qualifier) by adding two levels: one on top of it (metaterm) and one below it (resource type). This model shared with CISMeF was recently enhanced: a meta-term has now semantic links not only with MeSH terms but also with qualifiers and resource types.

To our knowledge, D'C is the first tool using the main advantage of the MeSH thesaurus (the explode command) to retrieve Internet Health resources. We even generalize this idea using our four level of thesauri: meta-term, MeSH term, qualifier and resource type. Therefore, a search in D'C can combine the explode command for keywords and qualifiers and also for meta-terms and resource types.

One main objective of D'C is to promote best medical practice and teaching. Therefore, the first prototype version of D'C includes high-quality documents available on the Internet on a priority basis. Although in the future, D'C will include all resources available in CISMeF, we plan to keep a core D'C composed only with high-quality documents creating a virtual encyclopedia.

D'C & CISMeF share a lot of hyperlinks because the browse (CISMeF) and search (D'C) strategies are complementary. At the bottom of a MeSH page of the "Browse CISMeF", the end-user can execute a simple search of D'C with the same MeSH term but with the explode command. On the other hand, from a search in D'C, the end-user can go to the "Browse CISMeF" on a MeSH page after clicking on the complete record.

D'C and CISMeF are mostly based on the MeSH thesaurus. Nonetheless, some of the features of UMLS [10] are used to enhance our structure model: (a) to create added "see also" relations in the browse CISMeF than those existing in the MeSH thesaurus and (b) to enhance the number of synonyms both in English and French. We do not yet use the main features of UMLS because: (a) after 5 years of CISMeF developments, we are using only 21% of the MeSH thesaurus and we use the manual "mapping" function in very rare cases; (b) most of the terms included in UMLS are not yet translated in to French. We think the "encapsulated" MeSH thesaurus used in D'C and CISMeF is already a good solution to catalog and index Internet Health resources.

Persons using D'C need some basic knowledge about the MeSH thesaurus. Since 1997, the RUH medical librarian is already giving a two hour training course about Medline and CISMeF. Since September 2000, he has added a ½ hour session about D'C.

The ultimate goal of D'C is to become the equivalent of Medline for cataloguing and indexing Internet Health resources. Therefore, D'C needs several improvements:

- A mapping function such as the one existing in the Internet Grateful Med Web site [http://igm.nlm.nih.gov/] to let inexperienced end-users type any word and translate it into neighboring MeSH terms.
- A "step by step" module such the one existing in the Ovid [http://www.ovid.com] and PubMed [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed] Web sites allowing more complex searches than the "Advanced search" module (e.g. including more than 4 MeSH terms or 4 meta-terms).
- In Doc'CISMeF vs. CISMeF, each keyword is no longer a MeSH Major Topic. Nonetheless, it is not yet possible in D'C to focus a search with MeSH Major Topic.
- D'C uses only floating qualifiers and not yet assigned qualifiers.

A search using meta-term(s) generates a pertinent request including all the "reserved" terms of the medical specialty (meta-term) without any knowledge of the MeSH thesaurus. This kind of search is very useful when a search using keywords gives few results. Then, the meta-term will allow the search to widen by including all the terms (key words, subheadings, resource type) of the medical speciality.

D'C is providing by default the broadest search which can be narrowed in further steps: e.g., an end-user types in the simple search "hospital" when needing the list of hospitals Web sites. This "reserved" term is in the D'C model a MeSH term and a resource type. Therefore D'C will select records indexed with the MeSH term or the resource type. To retrieve the list of French hospitals, the end-user has to choose the advanced search, selecting "hospital" in the resource typer field, and unselecting the explode command to avoid noise coming from resources indexed with resource types below hospital in the resource type tree.

Unlike PubMed, in the simple search, if a end-user types a MeSH term, the D’C displays results on resources indexed only with this term and not with a full text search because: (a) in D’C, we do not have an equivalent to PreMedline as all the resources included in D’C are indexed in the first stage and (b) we refused in the simple search to introduce noise with a complementary full-text search; in the advanced search, this is possible if the end-user chooses the “all fields” function which includes the MeSH term explosion and a full-text search.

We assume the simple search will mostly be used by Netizens and the advanced search by medical librarians and trained health professionals but further evaluation studies of its use will be needed to confirm this intuition.

D’C is also a MeSH “translator” through PubMed because D’C includes the F-MeSH (the MeSH in French) and the E-MeSH (MeSH in English) thesauri. If the end-user knows the F-MeSH, he can enter a F-MeSH term and D’C will generate automatically the equivalent PubMed request. In fact, to our knowledge, few users (some librarians) knows the F-MeSH. Therefore, this MeSH translator will be used “by chance” by end-users. The CISMeF team will focus on F-MeSH in the training sessions of D’C and will also emphasize the knowledge of the English version of the MeSH which is very important for the proper use of Medline. On the other hand, medical librarians who know the E-MeSH quite well will be able to perform their D’C searches in English.

Further challenges that D’C needs to address in the next few months are: (a) to expand its database, focusing on high-quality documents in evidence-base medicine, patient information and teaching material; for the latest, a subset version of D’C on that field is planned to be one of the main search tools of the French Medical University [11]; (b) to collaborate more closely with similar services, particularly in Europe (DDRT, HON and OMNI ; (c) to formally and directly assess how end-users use D’C. In the near future, we will measure, by questionnaire, its real usefulness for different communities (MDs, nurses, and patients) in the different French-speaking countries.

Conclusion

The aim of our study was to assist healthcare professionals and consumers to access more easily and accurately quality health information on the Internet. It is imperative that both catalogs and specialized research tools use thesaurus and metadata to describe and index useful resources.

Acknowledgements

D’C was supported in part by a grant from the Educnet programme (French Ministry of Research).

References

[1] Schatz BR. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science* 1997: 34 pp. 275-327.

[2] Flannery MR. Cataloging Internet resources. *Bull Med Libr Assoc* 1995: 83(2) pp. 211-5.

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

[4] Weibel S, Juha H. DC-5: The Helsinki Metadata Workshop; A Report on the Workshop and Subsequent Developments. D-Lib Magazine. 1998 February. Available from Internet: <http://www.dlib.org/dlib/february98/02weibel.html>.

[5] Thirion B, Darmoni SJ. Simplified access to MeSH Tree Structures on CISMeF. *Bull Med Libr Assoc* 1999: 87(4) pp. 480-1.

[6] Darmoni SJ, Thirion B. A standard metadata scheme for health resources. *J Am Med Inform Assoc* 2000: 7(1) pp.108-9.

[7] Centrale Santé. Net Scoring : critères de qualité de l'information de santé sur l'Internet 20 Apr 2000 [Web document, accessed 6 Oct 2000]. Available from Internet: <http://www.chu-rouen.fr/netscoring>.

[8] Darmoni SJ, Thirion B., Leroy JP, Douyère M., Piot J. The Use of Dublin Core Metadata in a Structured Health Resource Guide on the Internet. Mednet 2000, World Congress of the Internet in Medicine.

[9] Nelson SJ, Schopen M, Schulman JL, Arluk N. An Interlingual Database of MeSH Translations. 20 Aug. 2000 [Web document, accessed 6 Oct. 2000]. Available from Internet: http://www.nlm.nih.gov/mesh/intlmesh.html

[10] National Library of Medicine. Fact Sheet UMLS Metathesaurus. 12 Aug 1998 [Web document, accessed 6 Oct 2000]. Available from Internet: <http://www.nlm.nih.gov/pubs/factsheets/online_indexing_system.html>

[11] LeBeux P, Duff F, Fresnel A, Berland Y, Beuscart R, Burgun A, Brunetaud JM, Chatelier G, Darmoni SJ, Duvauferrier R, Fieschi P, Gillois P, Guille F, Kohler F, Pagonis D, Pouliquen B, Soula G, and Weber J. The French Virtual Medical University. In: Proceedings of MIE 2000, Sixteenth International Congress of the European Federation for Medical Informatics, 2000:554-61

Address for correspondence

Stefan J. Darmoni,
Rouen University Hospital, Computer and networks department,
1 rue de Germont F76031 Rouen Cedex, France
(Email: Stefan.Darmoni@chu-rouen.fr)