SDI Selecting, Describing, and Indexing: Did You Mean Automatically?

B. Thirion[1], J.P. Leroy[2], F. Baudic[1],
M. Douyère[2], J. Piot[1], S.J. Darmoni[2]

[1]Bibliothèque médicale, [2]Direction de l'Informatique et des Réseaux
Centre Hospitalier Universitaire,
76031 Rouen Cedex France
Fax: 02 32 88 87 86
E-mail: Benoit.Thirion@chu-rouen.fr

B. Thirion, librarian of the University Hospital Center of Rouen, France, is coordinator of CISMeF, the Catalogue and Index of French-language Medical Sites. The interest of this project lies especially in the joint effort to structure the information provided as well as in the use of the Dublin Core metadata.

Magaly Douyere, medical librarian at the University Hospital Center of Rouen, France, works on description, indexing, and maintenance of the database of CISMeF.

S. J. Darmoni, MD, PhD, is coordinator, along with B. Thirion, of the CISMeF project. The ISMeF research team currently has eight members. Darmoni is also the Advanced Technologies Manager, Computing and Networks Department, Rouen University Hospital. In addition, he is Associate Professor of Medical Informatics, Rouen Medical School (www.univ-rouen.fr/medecine/) and a member of the Perception System Information Lab

B. Thirion, J.P. Leroy, F. Baudic, M. Douyère, J. Piot, S.J. Darmoni. (2001). **SDI Selecting, Describing, and Indexing: Did You Mean Automatically?.** *Knowledge Organization, 28(3).* 137-140. 5 refs.

Introduction

Information available on the Internet is by definition accessible to all, for better or for worse. In the field of medicine, it is directed to health professionals together with their patients, and more rarely to a massive audience. But if there is a specialty that requires high quality information, it is certainly medicine.

A selection of information **prior to use** seems therefore indispensable. The automatic tools currently available are not satisfactory. A rigorous human selection made by information professionals (IP) with the advice of networks of experts, seems the most reasonable and efficient approach to gaining access to the incredible variety of this information. Distinct quality criteria may be applied, whether using a grid or not, together with a few hints of good practice occasionally suggested by common sense.

Resource description and indexing also seem to fall within the scope of IP's skills, because they are familiar with classification and indexing standards. Automatic indexing, although a very seductive technique in terms of cost and efficiency, does not yield convincing results. The "thinking being" is still far from finding this software equivalent. Considering the impressive scale of the task, it is becoming urgent to set up co-operative projects and to build sites that combine energies, the purpose of these sites being to select, describe and index the resources that can be found on the Internet. Shared cataloguing is more than ever a topical subject.

At the University Hospital Centre (CHU) in Rouen, we have been trying for four years to make our contribution to this initiative by creating CISMeF, the Catalogue and Index of French Medical sites [http://www.chu-rouen.fr/cismef]. This catalogue currently counts more than 6500 indexed sites and documents (May 1999). Our priorities for indexing are institutional sites and documents about factual medicine (consensus conferences and recommendations for good clinical practice) and teaching (online courses, multiple choice questions, case study, etc.). Information destined for patients is also one of our top priorities. CISMeF relies on the structure and key words of MeSH – the thesaurus available at the North American library of medicine (NLM), also used to create Medline, a bibliographical database. The French terms have been translated by INSERM (l'Institut National de la Santé et de la Recherche Médicale, the French National Institute for Health and Medical Research) and the bibliographical notes follow the recommendations related to the Dublin Core (DC) [http://purl.org/DC/index.htm]. Our metadata are also in line with recent developments made by DC (an example of which is given at the end of this article).

Within the framework of the «Information Gateway» proceedings, CISMeF obtained a «public interest experimentation» label in March 1998 – a label given by the French Interdepartmental Committee for Information Gateways and Services. Since November 1998 the CISMeF project has been considered as a first priority project financed by the University Agency of the French Speaking World (AUPELF-UREF) within the context of the French Speaking World Virtual University. CISMeF is a partner of CIDMEF (International Conference of French Speaking Faculty of Medicine Deans).

Finally, we aim at participating in the joint telemedicine and health technology incentive action led by the French Ministry of Education, Research and Technology for the creation of a Medical Virtual University.

We have also carried out a specific training action in the use of CISMeF for local patient associations.

SDI (Selecting, Describing, Indexing)

In a previous work (1) we had suggested that it was the duty of libraries to intervene in order to make an inventory of the resources available on the Internet (2). It should be noted though, that four years later the problem remains the same. In the field of medicine as in other fields of activity search engines such as AltaVista are invaluable tools for retrieving information about an institution or association. However, serious problems arise when the requested information is about a determined therapy or diagnosis. There is nothing surprising about this fact considering that the selection process does not rely on expertise.

High quality information in the field of medicine is increasingly being made available under the form of technical reports, recommendations for good clinical practice, consensus conferences, epidemiological data, image databanks, full text articles approved by reading committees, personal web sites of great informative value, etc. But this amount of information has to be validated and «acknowledged» useful by those who are its potential users, i.e. health and health information specialists and the patients.

Once this selection is performed, the information must be described in detail in order to be able to appreciate its usefulness beforehand. In this context, the search engine can only reflect the existing reality, that is to say a vague highlight of the first lines of page, free keywords and similar odd metadata. Some multidisciplinary directories like Nomade have improved the descriptive aspect of the data displayed but they have to face an ever-renewed difficulty in organizing resources from very diverse origins, which mingle indiscriminately second hand car dealers, hunters' associations and specific dictionaries dealing with health economy. Let us recall here that the MeSH thesaurus of medical terms, which will be the object of a separate paragraph, contains 19,232 specific terms distributed among nine hierarchical levels (1999 issue) and that a thorough exploration of its tree structure is necessary to enlarge or refine a search.

Indexing is the third point of our SDI triangle that must be carried out with standard and qualified tools. Automatic indexing is far from having shown its full potential in our specific domain; it may not be a coincidence that Medline, the National North American Library databank, uses the skills of some sixty full time employees to index the articles in 4300 journals included in this database. They do work with the help of software tools but the actual choice of the relevant key word is still in fact a human choice.

It would, of course, be totally impossible and overly ambitious to embrace the idea of indexing the whole of the Internet ... alone! But it would be equally unreasonable to expect the web surfer to rely exclusively on non-specialized search engines and directories to provide direct access to any useful online information. What is then the most relevant solution?

We think that before progress can be made on automation, it is necessary for IPs to set a specific organization for the creation of cooperative sites aiming at selecting, describing and indexing the resources of the Internet. Sites of this type like Sitebib [http://www.abf.asso.fr/sitebib/] (3) have already been created. This site is the result of a tight cooperation between web sites dedicated to library management and information science. Each of its members assumes a particular responsibility such as training, electronic newsletter or library site exploration.

Regarding medicine, the creation of CISMeF (French speaking world medical sites catalogue and index) was initiated in February 1995. To access an English language equivalent site web surfers can visit Omni [http://omni.ac.uk/], in which a particular effort has been made on selection, description and indexing; or HealthWeb [http://healthweb.org/], a site based on cooperation between various medical libraries.

Selecting

The text mining process, the first step of catalogue building, consists of browsing multidisciplinary directories like Nomade to extract any site related to health. A complementary search is then performed on sites that produce documents: ministries, knowledge associations and similar institutions. To complete this selection we rely on a number of major quality criteria originating from Netscoring classification [http:// www.chu-rouen.fr/dsii/publi/critqualv2.html] (4). We mention this classification because we were involved in its creation, but there exist many other works on quality criteria that make use of similar criteria including author's name or editing institution, last update for a document, etc. Some personal web sites can be taken into account when their informative content and the author's identity are clearly established. For example, a site on the training and education of diabetic patients was at first refused, then later accepted after a mail exchange with the author revealed that he was a hospital doctor whose qualifications are provided on his personal web page and that, in particular, the proposed online educational programme was the same programme proposed for diabetic patients in the hospital in which he worked. Revealing the author's identity is an important quality factor for a medical web site founded on the fact that a health specialist who agrees to disclose his identity also agrees to be subject to the judgement of other health specialists.

We often have to resort to a network of experts, created year after year, who are capable of validating the content of a web site and the quality of the information it contains.

Description

The establishment of information notices that describe the content of a site or document leaves no room for improvisation. Nevertheless, we did not wish to "reinvent the wheel" and desired to be in line with ongoing main institutional projects. This is why we chose to describe available resources using the DC format. Among its fifteen criteria (optional and reusable) we chose: author, date, description, editor, format, identifier, language, key words, title and resource type. As regards resource type, we joined the structuralists camp, and think that the resource types proposed by the DC project lack precision and therefore, cannot be used to describe medical resources correctly. We then decided to build our own online list of criteria [http://www.chu-rouen.fr/documed/typeressource.html], relying partly on Medline publication types that were extended to identify course, association, technical report, image databank, search structure and patient information.

Indexing and catalogue structure

The keywords used by CISMeF come from MeSH (Medical Subject Headings) specific terminology extracted from the NLM thesaurus that serves to index Medline references. Medline is the most widely used medical database in the world, which explains our choice. In fact, we think that just as DC will probably soon become a standard for resource description and the creation of metadata , MeSH terminology should be used to state online medical resources identity. Training in these standards is as applicable for web searchers as much it is for information producers. DC was intentionally made simple in order to allow authors to describe their own resources easily. This is also true for key words, but MeSH, with its 19,232 terms, presents a more complex approach than DC's fifteen criteria! For this reason it would be a great improvement if web site creators applied to experienced and competent IPs to create web sites or databases using a combination of the two standards: DC + MeSH.

CISMeF has not yet been organized dynamically and all pages are in html format. Information may be accessed through various indexes:

- An alphabetical index using the MeSH key-words
- A thematic index by biological and medical specialty which also provides access to CISMeF tree structure and metaterms (see below)
- A general index for all terms (equivalent to a one page swap index).
- A search engine allowing full text search on all pages of the catalogue. It is not a very technologically advanced search engine but should be replaced shortly by a tool that allows the use of Boolean operators and truncation.

MeSH tree structures are developed in French along with new MeSH keywords creation. The terms used are those translated by INSERM. Metaterms have been developed to access the MeSH tree structure more easily (5). From the key word «neurology», for example, it is possible to access any related subdomain such as: nervous system, neurosurgery, nervous and peripherical nervous system drugs, etc. We should keep in mind, though, that neophytes, those not familiar with Medline, as well as health professionals and patients, all use this catalogue. It is for this reason that we decided to set up training sessions in the use of CISMeF, designed for patient association leaders, who can then pass the knowledge on to their subscribers.

CISMeF is structured into five different levels:

1. Metaterms
2. Tree structures
3. Key-words (along with the «see also» or «not for coordination with» specific NLM recommendations)
4. Qualifiers (associated with keywords according to NLM rules)
5. Resource type

Perspectives

This catalogue is viewed every working day by 2500 machines for a total amount of 10,000 html pages (April 1999 figures). The email messages received regularly show a clear distinction between very varied categories of users such as health specialists, students and patients. CISMeF will most probably evolve to become a dynamic database, but we also wish to maintain its static structure that authorizes its access to computers that can read DC metadata.

Example of DC metadata use:

```
<meta name="DC.language" content="fre">
<meta name="DC.type" content=
    "(SCHEME=CISMeF)texte">
<meta name="DC.subject.keywords" content=
    "(SCHEME=MeSH)bases de données
    bibliographiques; databases, bibliographic">
<title>Bases de données bibliographiques : sites
    francophones</title>
<link rel="schema.mesh"
href"http://www.nlm.nih.gov/mesh/
    meshhome.html">
<link rel="schema.cismef"
href"http://www.chu-rouen.fr/documed/
    typeressource.html">
```

Bibliography

(1) Thirion B. Darmoni S.J. Les sites médicaux francophones sur Internet : le devoir d'ingérence des bibliothèques [http://www.enssib.fr/Enssib/bbf/bbf-98-3/08-THIRION.pdf]. Bulletin des Bibliothèques de France, 1998, n° 3, pp 42-5.

(2) Flannery MR. Cataloging Internet resources. Bull Med Libr Assoc 1995 Apr;83(2):211-5.

(3) Sigaud F. Sitebib : un exemple de coopération entre sites Web dans le domaine des bibliothèques et des sciences de l'information. In : L'Information scientifique et technique et l'outil Internet : expériences, recherches et enjeux pour les profesionnels de l'IST. Le Micro Bulletin Thématique LMB 76, 1999. Paris : CNRS.

(4) Darmoni S.J Leroux V. Daigne M. Thirion B. Santamia C. Duvaux C. Critères de qualité de l'information de santé sur l'Internet. In: Santé et Réseaux Informatiques, Informatique et Santé, 1998; 10: pp 162-174. [version mise à jour http://www.chu-rouen.fr/dsii/publi/critqualv2.html].

(5) Thirion B, Darmoni S.J. A simplified access to "MeSH Tree Structures " for health professionals and patients, our new "end-users" on the Internet. Bulletin of the Medical Library Association 1999; Oct (in press).