Organization of the Information about Health Resources on the Internet

Abstract: The growing of the World Wide Web and its unstructured nature makes searches difficult, and moreover the information published by web sites is not always guaranteed. Usual searching engines are not enough to satisfy the needs of users and ease access to credible information. This paper focuses on the design of a medical information retrieval system that uses the MeSH controlled language and the fields structured to organize the information, mostly from Dublin Core metadata format. We use a set of criteria to guarantee the minimum level of information quality and credibility. The objective of such a system is to provide an efficient mechanism of information retrieval with greater searching capabilities than other subject gateways or search engines on account of its structure.

1. Introduction

An unquestionable fact is the massive and continuous growth of the information resources available on Internet, and especially on the World Wide Web. The facility to publish and the lack of control of the information are factors that have favoured the explosion of information on Internet which has become the greatest store of world-wide knowledge and, with no doubt, in the most used information sources, mainly being free. Nevertheless, the information that we can find can be of many kinds and created for multiple purposes, existing in different degrees of quality and reliability, due to the lack of screening and control.

These previous lines summarize the informative conditions of Internet: excess of information and questionable quality. This situation is generalized in any scope and discipline, but in the field of the health it acquires a great importance specially for two reasons; first of all, it is probably the field where most information is published; secondly, it has been demonstrated that the misinformation can be harmful for patients who use Internet as an alternative to find information on the treatment of their diseases (Kiley et al., 2002), and that although this information even comes from reliable and credible resources can not be correct for certain kinds of users (Impicciatore et al., 1997).

Under this situation, healthcare professionals and users face new challenges to locate useful, reliable and quality information. In order to try to solve the quality
aspect, different initiatives have arisen to develop ethical norms and to regulate the information quality, although it has been demonstrated that there are important gaps that must be treated in future (Risk, 2001). The main difficulties to evaluate the information quality are derived from the complex and multidimensional nature, while subjective, of the very concept of "information quality", and from the developed mechanisms and instruments to carry out the evaluation. Although different review sites have developed and used instruments to measure the information quality, it has not been possible to verify the construct validity and the interobserver reliability and its systems of rating are incomplete and of questionable utility (Jadad, 1998; Gagliardi, 2002).

In addition to the quality, the other important problem is how to search and locate relevant information. Although there is a great amount of directories and search engines, not only general but also specialized in health subjects, they lack of a clear and organized presentation of the information adapted to medicine and health world, which limits the utility of these tools in spite of the great number of resources that can contain. In order to solve this problem, several gateways of quality medical resources on Internet have arisen in the last years among them it is possible to emphasize in English language OMNI, Medical Matrix, Health on the Net, CliniWeb, Healthfinder, etc., and in French language CISMeF.

The indexing and organization of the resources in these subject gateways is based, generally, in the use of the controlled language MeSH (Medical Subject Headings) of the U.S. National Library of Medicine. Nevertheless, this is not enough to search for and to efficiently retrieve the information. Berland et al. show that the access to health information using search engines and simple search terms is not efficient (Berland et al., 2001), and with the exception of CISMeF, the possibilities of retrieval by different fields in the data bases of these gateways are very limited.

2. Objectives

The objective of this paper is to show an alternative proposal to the existing search systems and to describe INDISALUD (Index of Resources of the Health on Internet in Spanish), an Information Retrieval System (IRS) developed in the frame of a research project granted by the Vicerrectorado de Investigación of the Universidad de Zaragoza with the following purposes:

To structure, organize, describe, classify and index the health web resources in Spanish in a standardized and suitable way.

To expand and improve the retrieval options offered by the search services in the Web.

To provide enough information so that the professional and end-user can evaluate quickly his pertinence and some aspects related to quality.

3. Methods

To create and design the Information Retrieval System we have considered the following aspects:

a) Definition of the resources selection and inclusion criteria, in such a way that they act as a filter and they guarantee his potential relevance and credibility. In addition to the subject scope and resources language, criteria of inclusion and
exclusion have been defined related to the information reliability, adapted from the ones used in OMNI (BIOME, 2001), whose observance is mandatory to add the resources to the database.

b) To define the database structure: the information is organized and structured in 22 fields, 15 of them corresponding with the Dublin Core metadata. In Table I the fields are described; to assure the information consistency, some fields take their values from controlled lists.

c) To create a database with the resources candidates to be introduced in the INDISALUD system obtained by a search in several engines not only general but also specific, asking for the terms in Spanish language “salud” (health) or “medicina” (medicine). Once the URL addresses have been obtained, the following step is to identify the main address and to eliminate the repeated resources or those that belong to the same site. The resources thus obtained are those that have been verified with the inclusion criteria. Finally, this list closed when the 2,000 valid resources have been reached, the amount with which the service will be started initially.

d) The system input consists of the resource analysis and description according to the structure presented in Table I. The most relevant aspects correspond to the subject indexing and the abstracting, tasks developed traditionally by documentation professionals with the mission to represent the content of documents to be able to retrieve them later.

The subject indexing is made using as much descriptors of MeSH controlled language as terms of natural language, which facilitates the content representation in a consistent way and the exhaustive and accuracy retrieval of information. The fields in which the thematic subject of the resources is indexed has been organized hierarchically in several levels: medical specialty, descriptors, subheadings, identifiers and resource type.

These input tasks are developed by students of Librarianship and Documentation previously trained not only in the Web resources description and evaluation but also in techniques of indexing and abstracting and in the use of the MeSH controlled vocabulary. Later, the introduced resources are reviewed by documentation and medicine professionals, that are also in charge of the translation of the MeSH headings and subheadings into Spanish, avoiding therefore the possible variations in the translation and guaranteeing their consistency.

e) Retrieval mechanism: the search and location of resources can be made in two different ways:

- A browsing system through an alphabetical classification and a thematic classification of the resources, structured in five hierarchic levels: category of specialties, MeSH descriptors, subheadings, identifiers and resource type, which allows us to expand or to reduce the search by anyone of these levels.

- Search interface: its design allows us to enter the search terms directly, being we able to make the search in all the database fields or limiting it to certain fields, and to select the terms by consulting the index files. The search features include the use of Boolean operators, proximity operators, truncating, rank of dates, and explosion of terms in the fields of descriptors and resource types.
<table>
<thead>
<tr>
<th><strong>Campo</strong></th>
<th><strong>Descripción</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Resource name</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Location (URL) Address</td>
</tr>
<tr>
<td>Author</td>
<td>Author/s or person/s responsible of the resource intellectual content. It can be a person, organization or service.</td>
</tr>
<tr>
<td>Organization</td>
<td>Author's institution or organization, including the department (if exists).</td>
</tr>
<tr>
<td>Editor</td>
<td>The responsible entity for making the resource available on the Net in its present format.</td>
</tr>
<tr>
<td>Country</td>
<td>Country where the resource is. Controlled list.</td>
</tr>
<tr>
<td>Province or State</td>
<td>Province or State where the resource is.</td>
</tr>
<tr>
<td>City</td>
<td>City where the resource is.</td>
</tr>
<tr>
<td>E-mail</td>
<td>Author e-mail of resource or publisher.</td>
</tr>
<tr>
<td>Abstract</td>
<td>Descriptive summary of the resource content.</td>
</tr>
<tr>
<td>Specialty</td>
<td>Medical specialty or specialties to which the resource belongs. Controlled list.</td>
</tr>
<tr>
<td>Descriptors</td>
<td>Subject indexing with the MeSH descriptors. The most specific descriptors will be used. In case it is required, MeSH subheadings will be included, added to descriptors. Controlled list.</td>
</tr>
<tr>
<td>Identifiers</td>
<td>Descriptors or identifiers of the resource not included in the MeSH. They will be used to describe the content in case there are not precise MeSH descriptors.</td>
</tr>
<tr>
<td>Resource Type</td>
<td>Type, nature or category of the resource. Controlled list.</td>
</tr>
<tr>
<td>Format</td>
<td>The physical or digital manifestation of the resource. Controlled list.</td>
</tr>
<tr>
<td>Source</td>
<td>Reference from where the resource is derived.</td>
</tr>
<tr>
<td>Creation date</td>
<td>Date when the resource in its present format has been created.</td>
</tr>
<tr>
<td>Update date</td>
<td>Last date of update of the resource.</td>
</tr>
<tr>
<td>Consultation date</td>
<td>Date in which the resource has been consulted.</td>
</tr>
<tr>
<td>Check date</td>
<td>Date in which the resource has been reviewed for the last time.</td>
</tr>
<tr>
<td>Audience</td>
<td>To whom the resource is directed, professionals or public in general.</td>
</tr>
<tr>
<td>Status</td>
<td>Active or inactive.</td>
</tr>
</tbody>
</table>

Table 1. INDISALUD database structure

1) Results display: the retrieved resources through the search interface show up sorted by update date or creation date, and it is possible to display them in two kinds of formats:

- Short format: it includes the title, URL, author, abstract and descriptor fields.
4. Results and Discussion

The result is an Information Retrieval System with Spanish web resources related to health aimed not only to the healthcare professionals but also to the public in general. Its main characteristics lie in the conceptual model of the information structure, the added value information from the abstract and the indexing using the most used controlled language in the medicine field, and the different search possibilities in ever field with the consequent advantages of improving both recall and precision. All it conforms an efficient Information Retrieval System that differs more or less with the existing search engines and subject gateways; related to the first one the differences are derived specially as far as offering a structured right information and to the fact that the indization is done by people with a medicine specific controlled vocabulary; related to the second one, because almost always they only offer a box to enter the search terms and they lack of options like search for in all the fields or to use the terms.

Another important utility of INDISALUD is that the user has access only to reliable and credible resources since they must fulfill inclusion criteria that act as a filter. The lack of a validated and reliable evaluation instrument, as well as low values in the indices of concordance between the judgments made by medical experts in the evaluation of the resources quality in the Web (Craigie et al, 2001), and the thought that the last judge in assessing the information usefulness and quality is the end user, has made us desist from carrying out an evaluation of the resources quality. Nevertheless, with the structure proposed we contemplate certain dimensions and characteristics related to the information quality that can be evaluated quickly by the user:

- Fast identification of reliable and credible resources by means of "author", "organization" and "editor" fields, that they provide information on the authorship and authority criteria of the origin of the resource, and the "source" field for those resources that have been published in printed medium and that have gone already through a review process.
- Selection of resources aimed at a determined audience through the "audience" field, in such a way that in the search those resources directed to professionals or interesting only for patients or users in general can be discriminated.
- Identifying the currently and update of the information through the "creation date" and "update date" fields.
- Selecting and evaluating the potential relevance or pertinence of the resources through the subject fields like "title", "abstract", "speciality", "descriptors" and "identifiers", as well as from "resource type".

INDISALUD will become accessible to the public in September 2002. From this date resources will be added to the system coming from both the periodic search in search engines and the sending by webmasters. The challenge is to
continue growing in the number of offered resources, but especially to maintain the
database updated. Verifying the operativity of the resources periodically is simple
and automatic; more effort implies to verify that a resource has been updated, since
it will not only affect the update date but very specially the thematic fields like title,
summary and descriptors that in case it is necessary will have to be modified to
maintain their informative value and to reflect the resource accurately.

5. Conclusions

The great amount of existing information on Internet, the different types of
formats, the lack of structure and the variable degree of quality are some of the
factors that cause the valid and useful information search in this environment to be
difficult. As happened fifty years ago with the phenomenon of the scientific
information explosion that gave rise to the appearance of the online systems with
their referential or full-text databases, it becomes necessary at the present time to
develop information retrieval systems, called thematic gateways or otherwise, that
allow the search and efficient retrieval of relevant information and with certain
quality and credibility assurance on Internet.

The efficient retrieval depends not only on the capacities of the search
system and the searcher ability, but also on how the information is organized and
structured and on the precision whereupon the resource content is represented
during the description, indexing and abstracting process.

The structured information model proposed in this paper does not represent
any new development in relation to the traditional bibliographical databases. It
simply tries to know how to take advantage of the accumulated experience and
knowledge throughout the years and to implement those characteristics of the
traditional systems that still at the present time represents an advantage on the
current systems, or in form of structure, metadata or retrieval languages.

Notes

1 A search made in Google (http://www.google.com) the 8th of March 2002 with the simple
term "health" recovered 68.500.000 resources.

References

Berland, G.K.; Elliott, M.N.; Morales, L.S.; Algazy, J.I.; Kravitz, R.L.; Broder, M.S.;
Kanouse, D.E.; Muñoz, J.A.; Puyol, J.A.; Lara, M.; Watkins, K.E.; Yang, H.; McGlynn,
E.A. (2001). Health information on the Internet: accessibility, quality, and readability in

paper on gateway specific quality issues. Quality issues for OMNI and NMAP.*

on the Internet: An examination of expert’s ratings. *Journal of Medical Internet
02/8/2002.

Gagliardi, A.G.; Jadad, A.R.; (2002). Examination of instruments used to rate quality of
health information on the internet: chronicle of a voyage with an unclear destination.

