A Preliminary Trial of Tagging On-line Documents Using "Medical Core Metadata"

Yang Gong M.D. M.S., William Hersh M.D.
Division of Medical Informatics and Outcomes Research
Oregon Health Sciences University(OHSU), Portland, Oregon

Introduction

Finding documents on the World Wide Web relevant to a specific medical information need can be difficult. To address this problem, we have developed a set of document content description tags, or *metadata encodings*, which can be used to promote disciplined search access to Internet medical documents.

Medical Core Metadata (MCM) is a set of metadata elements designed for Web-based biomedical documents, [1], based on the Dublin Core Metadata standard (Table 1). [2] Documents tagged with MCM markup text can be retrieved by search engines that recognize its fields. The goal of MCM is to aid clinicians, researchers, and healthcare consumers in the retrieval of useful documents from the Internet and facilitate document retrieval by search engines. The purpose of this trial was to assess what percentage of MCM's 15 fields could be filled and what kinds of problems were encountered during the indexing.

Methods

A total of 100 URLs were randomly selected from CliniWeb database, [3] which contains the Uniform Resource Locators (URLs) of nearly 10,000 URLs, all of which are medically-oriented Web pages written at the level of health care professionals. We attempted to identify which of the 15 fields of MCM could be tagged from information available from the Web page or its site.

DC.title
DC.creator
DC.subject
DC.description
DC.publisher
DC.date
DC.contributor
DC.type
DC.format
DC.identifier
DC.source
DC.language
DC.relation
DC.coverage
DC.rights

Table 1 - Dublin Core Metadata elements.

Results

When the URLs in the CliniWeb database were revisited, we found that 4% had been changed and were no longer available. Some of the fields could be tagged in virtually all instances, such as title (100%), creator (100%), subject (100%), format (96%), language (96%) and type (96%). This is because most pages contained this basic information or it could be easily found by browsing the Web site. At the other end of the spectrum, there were some metadata pages for which the tags could not be found the majority of the time, such as relation (30%), contributor (30%), and source (29%). In the middle, the tagging percentage from the fields of coverage, rights and date were 70%, 82%, and 83% respectively.

Some fields were filled with identical information for all 100 URLs. All language fields were filled with English because CliniWeb contains only English Web pages. Similarly, since all of the pages were HTML Web pages, the format fields were all tagged with text/html.

For the type field, we used an enumerated list and added elements as necessary, so all pages could be tagged. We ended up using the following elements: textbook, encyclopedia, manual, chart, bulletin, profile, report, photo, pathological section, fact sheet, guideline, table, and x-ray.

Conclusions

Most of the 15 MCM fields can be tagged with the information obtained by browsing the Web pages, or consulting the upper level of the site. Additional information will be required to index all of the MCM fields for all Web pages.

References

[1] Malet G, et al., A model for enhancing Internet medical document retrieval with "medical core metadata". Journal of the American Medical Informatics Association, 1999. 6: 183-208.

[2] Hersh WR, et al., CliniWeb: managing clinical information on the World Wide Web. Journal of the American Medical Informatics Association, 1996. 3: 273-280.

[3] Weibel SL and Koch T, The Dublin Core Metadata Initiative: mission, current activities, and future directions. D-Lib Magazine, 2000. 6(12), http://www.dlib.org/dlib/december00/weibel/12weibel.html.

1067—5027/01/$5.00