Model Formulation ■

A Model for Enhancing Internet Medical Document Retrieval with "Medical Core Metadata"

GARY MALET, DO, FELIX MUNOZ, RICHARD APPLEYARD, PHD, WILLIAM HERSH, MD

Abstract **Objective:** Finding documents on the World Wide Web relevant to a specific medical information need can be difficult. The goal of this work is to define a set of document content description tags, or *metadata encodings*, that can be used to promote disciplined search access to Internet medical documents.

Design: The authors based their approach on a proposed metadata standard, the Dublin Core Metadata Element Set, which has recently been submitted to the Internet Engineering Task Force. Their model also incorporates the National Library of Medicine's Medical Subject Headings (MeSH) vocabulary and MEDLINE-type content descriptions.

Results: The model defines a medical core metadata set that can be used to describe the metadata for a wide variety of Internet documents.

Conclusions: The authors propose that their medical core metadata set be used to assign metadata to medical documents to facilitate document retrieval by Internet search engines.

■ **JAMIA.** 1999;6:163–172.

The wealth of resources available on the Internet has stimulated information scientists to consider new models for knowledge retrieval. In theory, the Internet's browsable, searchable, and hyperlinked interface should improve the speed and ease with which users obtain relevant materials. Authors could offer intuitive connections from their documents to remote sites and place their documents in the context of existing literature. Clinical case experiences could be simulated more successfully using the Internet's multimedia features. Despite these inherent advantages, the World Wide Web has established a reputation as a substandard source of health care information that is inadequate to serve the information needs of health care providers.[1,2] A recent study indicates that only 46 percent of surveyed physicians agree that the Internet is a source of timely, accurate, relevant, and objective content.[3]

Use of information resources that are accessible over the Internet for information retrieval in the medical sciences presents a number of challenges. Foremost among these are more reliable navigation tools, search utilities, and filters for content and quality.

In the past, information retrieval (IR) from large medical databases has been assisted by keyword matches on well-constructed resource descriptions, or *metadata*. Sophisticated approaches to defining the syntax and semantics for these content tags have evolved among professional indexers. For example, MEDLINE indexers tag articles with metadata, each tag containing a description of one the article's characteristics, such as author name, title, or journal name. The searchability of MEDLINE documents is also enhanced by another form of metadata, which are terms from the Medical Subject Headings (MeSH) vocabulary that describe an article's content.[4]

The use of MeSH makes it possible to retrieve articles by matching the subject descriptions assigned by the indexer, without the subject terms necessarily being in the article. The MEDLINE environment also provides a

Affiliation of the authors: Oregon Health Sciences University, Portland, Oregon.

This work was supported by Fellowship Training grant LM07088 from the National Library of Medicine.

Correspondence and reprints: Gary Malet, DO, 3181 SW Sam Jackson Park Road, Portland, OR 97201;
e-mail: ⟨gmalet@teleport.com⟩.

Received for publication: 4/1/98; accepted for publication: 10/23/98.

quality filter, since the high quality of the material is ensured in advance of indexing by the journal selection and peer review processes.

Compared with MEDLINE, the Web offers a more heterogeneous and dynamic environment for information retrieval. There has been no standard for content descriptions for subject coverage, type of resource, or relationship to other documents. As a result, the main approach to searching is through the use of global search engines, such as, Alta Vista (http://altavista.digital.com/), Excite (http://www.excite.com/), and Infoseek (http://www.infoseek.com/), which provide access to only a limited number of documents and uncontrolled text words in the documents. In addition, there is no widely accepted quality control process for Web-based documents.

The purpose of this paper is to propose a standard metadata schema for health and medicine resources. We introduce a metadata syntax and semantics, compatible with HTML code, that allows Web medical authors to tag their documents for more effective retrieval. First, we cover the principles of metadata tagging on the Internet. Second, we introduce the Dublin Core initiative, a proposed metadata schema for Web documents. Third, we present our metadata syntax and semantics for Web medical content, which are compliant with both the Dublin Core effort and the HTML specifications. Finally, we discuss how information retrieval may be implemented with medical core metadata.

Metadata Tagging on the Internet: The HTML ⟨META⟩ Tag

Web browsers interpret the HTML of Web documents and display appropriate font or images. The ⟨META⟩ tag in HTML was designed to contain any information about the document that the author deemed relevant but not required to be displayed. The HTML 2.0–3.2 versions of this optional tag, which can be implemented within the ⟨HEAD⟩ . . . ⟨/HEAD⟩ section of a Web page's code, may contain the attributes NAME and CONTENT. This gives the tag an attribute–value pair structure, with NAME containing the name of the *attribute* of the document and CONTENT containing the *value* assigned to that attribute.

This attribute–value pair provides a mechanism to describe the properties of a document (e.g., author, expiration date, a list of key words) and values assigned to those properties. Many Web developers have already made use of the ⟨META⟩ tag to describe their documents. However, without an agreement regarding content descriptions and their syntax and semantics, search and retrieval utilities can only provide word matches for META tag fields. For example, when assigning the author's name to a Web page, Developer A may have used *Author* as the NAME of the document descriptor while Developer B may have used *Creator*. Or when Developer A assigned a value for *Date*, he or she may have written it as *1998-1-1* while Developer B may have written it as *January 1st, 1998*. Hence, there are two main problems with the current ⟨META⟩ tag syntax: There is not a standard naming schema for the document descriptors, and the syntax does not allow for standardized value descriptors, e.g., it does not allow a developer to define a value for *Date* written in specific standard format.

Defining a Core Metadata Syntax and Semantics: The Dublin Core Metadata Initiative

The information retrieval problem of defining standard metadata content descriptions has brought together a cross-disciplinary collection of information scientists, librarians, and interested others to form the Dublin Core (DC) Initiative (http://www.purl.org/metadata/dublin_core/). This group, spearheaded by the Online Computer Library Center, Inc. (Dublin, Ohio), has established an international consensus on the syntax and semantics of Internet-based content descriptions through a series of workshops. They have set forth a core metadata element set (http://www.roads.lut.ac.uk/lists/meta2/1998/02/0061.html) that has been submitted as a formal request for comment to the Internet Engineering Task Force.

In its most basic form, a DC-compliant metadata tag follows the standard HTML ⟨META⟩ tag syntax:

```
⟨META NAME = "DC.element_name" CONTENT
                                   = "element_value"⟩
```

For the NAME value, the DC introduces the concept of a *scheme identifier*. The scheme identifier in this case is DC, which defines the DC as the entity that holds the definition for the scheme used. The use of this syntax is straightforward. In order to define the author of a document, for example, one would write the tag as follows:

```
⟨META NAME = "DC.creator" CONTENT
                                   = "Gary Malet"⟩
```

Any developer trying to determine the meaning of the element name, in this case *creator*, needs only look at the DC's definition of this element.

Table 1 ■

The Dublin Core Metadata Element Set

Dublin Core Type	Definition
Dc.title	The name given to the resource
DC.creator	The person or organization primarily responsible for creating the intellectual content of the resource
DC.subject	The topic of the resource
DC.description	A textual description of the content of the resource
DC.publisher	The entity responsible for making the resource available in its present form
DC.date	A date associated with the creation or availability of the resource
DC.contributor	A person or organization not specified in a creator element who has made a significant intellectual contribution to the resource but whose contribution is secondary to any person or organization specified in a creator element
DC.type	The category of the resource
DC.format	The data format of the resource, used to identify the software and possibly hardware that might be needed to display or operate the resource
DC.identifier	A string or number used to uniquely identify the resource
DC.source	Information about a second resource from which the present resource is derived
DC.language	The language of the intellectual content of the resource
DC.relation	An identifier of a second resource and its relationship to the present resource
DC.coverage	The spatial or temporal characteristics of the intellectual content of the resource
DC.rights	A rights management statement, an identifier that links to a rights management statement, or an identifier that links to a service providing information about rights management for the resource

The DC has proposed a set of 15 basic elements to provide resource descriptions for a Web document (Table 1). In many cases, however, these basic elements are not sufficient to define all the metadata a developer wants to assign to a document. For example, none of the elements can be used semantically to assign the author's e-mail address to a document. To solve this situation, the DC Working Group has proposed the use of Dublin core qualifiers. TYPE and SCHEME are two of such qualifiers. (The SCHEME qualifier is distinct from the scheme identifier introduced above.)

The DC qualifier TYPE allows the refining of the definition of the 15 core elements. The DC syntax for the use of TYPE is:

```
⟨META NAME = ″DC.element_name.TYPE_
        identifier″ CONTENT = ″element_value″⟩
```

For example, if one needs to assign metadata for the author's e-mail address, one can subtype the *creator* element, to refine the definition of its value:

```
⟨META NAME = ″DC.creator.email″ CONTENT
                        = ″maletg@ohsu.edu″⟩
```

The SCHEME qualifier, on the other hand, is used to interpret the value of the content. The proposed use of this qualifier is to facilitate the assignment of standardized values to the metadata tags, introducing some degree of consistency among values assigned by developers. This standardization can take two main forms: the use of a standard format for the value assigned to a metadata tag, or the use of a standard vocabulary to assign a value to a metadata tag, or both. The syntax of a DC metadata element containing a SCHEME qualifier (in HTML 3.2) is as follows:

```
⟨META NAME = "DC.element_name" CONTENT
        = "(SCHEME=identifier) element_value"⟩
```

Notice that the use of quotation marks around the scheme value is not allowed. This is very important for the correct interpretation of the CONTENT value by browsers, Web crawlers, and other tools.

This syntax facilitates the need to assign a scheme to the CONTENT value. For example, when assigning the CONTENT value for the DC *Format* element, one could write *Web page* or *HTML*. This is accepted as DC-compliant. These values, however, are not standard, and a Web crawler trying to interpret this value could be at a loss. If one decided to standardize this value, however, one could use the existing Internet media types (IMT, also known as multipurpose internet mail extensions, or MIME, types) to describe the document, as follows:

```
⟨META NAME = "DC.format" CONTENT
        = "(SCHEME=imt) text/html"⟩
```

Looking at this tag, a Web crawler knows that the CONTENT value that follows is an IMT and that it follows the IMT syntax. Standardization of the CONTENT values would allow, among other things, automation of the interpretation of these values.

To facilitate both the semantic understanding and the interpretation of the CONTENT value, the DC proposes the use of the existing HTML ⟨LINK⟩ tag, which would have the function of defining a URL reference for the scheme used for both the TYPE and the SCHEME. For example, the following tag defines the URL of the DC document that explains the semantics of the *format* type:

```
⟨LINK REL = SCHEMA.dc HREF = "http://purl.org/
        metadata/dublin_core_elements#format"⟩
```

The following tag defines the URL of the IMT page, which explains the value of *text/html* given to the *format* type:

```
⟨LINK REL = SCHEMA.imt HREF
    = "http://sunsite.auc.dk/RFC/rfc/rfc2046.html"⟩
```

The use of the ⟨LINK⟩ tag is recommended but not required. However, if the tag is used, it must be placed immediately after the ⟨META⟩ tag that it helps define. For example:

```
⟨META NAME = "DC.format" CONTENT
        = "(SCHEME=imt) text/html"⟩

⟨LINK REL=SCHEMA.imt HREF
    = "http://sunsite.auc.dk/RFC/rfc/rfc2046.html"⟩
```

Recently, the proposed HTML 4.0 specification has been approved, although as of this writing (August 1998) it is not used in any major commercial browsers. Of interest to us is that this HTML version has implemented proposed Dublin Core qualifiers SCHEME as an attribute of the ⟨META⟩ tag. What this means is that, to define a scheme for the content value, the Web developer no longer has to add the scheme's name inside the value of the CONTENT attribute but can add it as separate attribute, as follows:

```
⟨META NAME = "element_name" SCHEME
    ="scheme_name" CONTENT = "element_value"⟩
```

Notice that the use of quotation marks around the scheme name is now allowed.

This has clear benefits for developing tools that make use of the ⟨META⟩ tag. In the HTML 2.0–3.2 syntax the element value is a complex of both the scheme construct and the underlying element value. Tools need to separate the two before going into any tasks to use the element value. With the HTML 4.0 syntax this extra step is no longer necessary. In HTML 2.0–3.2, for example:

```
⟨META NAME = "DC.subject" CONTENT
        ="(SCHEME=MCM-MeSHTerm)
            *Myocarditis/diagnosis, drug therapy"⟩
```

whereas in HTML 4.0:

```
⟨META NAME = "DC.subject" SCHEME
        ="MCM-MeSHTerm" CONTENT
        ="*Myocarditis/diagnosis, drug therapy"⟩
```

The medical core metadata syntax and semantics can be translated to HTML 4.0 or XML documents using translation tools. The W3C Resource Description Framework has proposed standards by which vocabularies and metadata semantics may be defined by a particular resource description community such as medicine. A public version of the W3C schema draft specification is available at http://www.w3.org/TR/WD-RDF-schema.

Extending the Core Metadata Approach for the Medical Knowledge Domain

The Medical Core Metadata (MCM) project has developed a set of content descriptions for the medical domain that builds on the metadata set offered by the Dublin Core Metadata working group. We have adopted the syntax and semantics for the metadata elements from the DC effort and extended these with refinements for medical resource types and the use of controlled language subject descriptions. We derived our list of metadata rules and protocols from those applied to documents in MEDLINE records—in particular, MeSH subject headings and publication types. We have defined new content descriptions to handle Internet resources such as images in an autopsy database or clinical discussion forums. We offer one approach to subject term assignment, the MCM-MeSHTerm scheme, and another for resource type assignment, the MCM-ResourceType scheme.

The MCM-MeSHTerm Scheme

The DC element Subject (DC.subject) exploits discipline-based controlled languages. Fortunately, medicine has a well-developed controlled vocabulary to aid information retrieval systems—the National Library of Medicine's MeSH thesaurus. Indeed, MEDLINE has achieved its prominence, in part, because of its effective selection and assignment of MeSH terms. Following the procedures used in indexing MEDLINE,[5] we propose that Web developers use the most specific MeSH terms available to describe the subject content of their documents.

One issue in using MeSH terms for the metadata subject is how to represent them in the HTML file. MeSH terms can be represented in three different ways:

- *The character string of the term*. The advantage of the string is its readability by humans. Its disadvantage is that direct user entry or editing of the HTML code may render the term unrecognizable by indexing programs that expect the exact form. In addition, the string itself may not be persistent (i.e., it may change as the vernacular changes).
- *The tree address*. These identifiers have no real advantages and several disadvantages, such as nonpermanence (e.g., tree addresses can change) and multiple representations for a single term (i.e., so terms can reside in more than one hierarchy).
- *The MeSH unique identifier* (labeled the DUI in the UMLS Metathesaurus). The advantage of the unique identifier is that it is persistent, even when the string of the term changes, and is unlikely to be entered or edited by users, thus minimizing chances for errors in the HTML code.

We have chosen to utilize both the character string and the unique identifier in the MCM system. This preserves the readability of the string but allows the easier computability of the unique identifier code. The string and the unique identifier will be separated by the pound sign (#), thus allowing most search engines to discern a break between the two. This approach will necessitate the use of metadata editing tools, but we believe this will be beneficial, since manual entry of editing of MeSH strings is likely to result in errors. The use of the unique identifier will allow for changes in the string over time, and editing tools should be able to update sites periodically for the small number of MeSH terms whose strings change with each annual edition of the vocabulary. Also, the use of unique identifiers will allow the coding of the subject with MeSH terms in other languages while allowing crawlers to map the subject to another language.

The ⟨META⟩ tag that makes use of this scheme will have as a value for the CONTENT attribute a MeSH term containing the string and unique identifier of the term separated by a pound sign. For example, the following is the subject metadata tag for a document that has *myocarditis* as a subject:

```
⟨META NAME = "DC.subject" CONTENT
        = "(SCHEME=MCM-MeSHTerm)
                    Myocarditis#UI009205"⟩
```

The MESH Subheadings and Major Subject Schemes

When cataloging documents, MEDLINE makes use of MeSH subheadings to focus the context of the subject heading assigned to a document. The MeSH subheadings can be implemented into the MCM syntax that we have introduced. For example, if the subject of the document were *the diagnosis and drug therapy of myocarditis*, no MeSH term would be able to convey this. However, by making use of the subheadings, the subject can be expressed as follows:

```
⟨META NAME = "DC.subject" CONTENT
    = "(SCHEME=MCM-MeSHTerm)Myocarditis/
        diagnosis, drug therapy#UI009205/DI,DT"⟩
```

Some implementations of MEDLINE define a "Major Subject" content description for documents. This tag

Table 2 ■

Medical Core Metadata Resource Types

MCM Resource Type	MCM Metadata Syntax	Medline Publication Type	Definition
Meeting	⟨META NAME = “DC.type” CONTENT = ″(SCHEME=MCM-ResourceType)Clinical Conference”⟩	Addresses, consensus conferences, lectures, meeting reports	Meeting announcements and reports
Directory	⟨META NAME = “DC.type” CONTENT = ″(SCHEME=MCM-ResourceType)Directory”⟩	Directory, periodical index	A list of items from other sources
Abstract	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Abstract”⟩	None	Introduction to the content of full text articles or resources
Homepage	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Homepages”⟩	None	Institutional or personal resource starting points
News	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)News”⟩	News	Releases, newsletters, and updates
Cases	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Cases”⟩	None	Case presentations
Images	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Image”⟩	None	Pathology, radiographic, and clinical images
Review	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Review”⟩	Review-academic, review-tutorial, etc.	Analysis of research reports or synopses of referenced materials
Study	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Study”⟩	Clinical trials, metanalysis	Formal, peer-reviewed, structured, referenced research reports and clinical trials
Procedure	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Procedure”⟩	Technical report	Interventions, techniques, surgeries, instrumentation, technical manuals
Educational material	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Educational Material”⟩	Bibliography, comments, editorials, letters, retracted publication,	Default category that includes learning modules, lectures, forms, continuing medical education materials, brief items, tables, charts, tracings, and algorithms
Video	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Video”⟩	None	Video transmissions or clips
Audio	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Procedures”⟩	None	Sound clips, radio programs
Database	⟨META NAME = “DC.type” CONTENT = “(SCHEME=MCM-ResourceType)Database”⟩	None	Searchable collection of items or documents

Table 2 ■

Medical Core Metadata Resource Types, *continued*

MCM Resource Type	MCM Metadata Syntax	MEDLINE Publication Type	Definition
Textbook	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)Textbook"⟩	Dictionary	Includes encyclopedia and sections of textbooks
FAQ	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)FAQ"⟩	Guidelines, monographs	Reference sources that include subject-specific instructions, tutorials, standards, protocols, or critical paths
Software	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)Forums"⟩	None	Decision tools, interfaces for queries of formulas, and computer programs
Patient education	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)Forums"⟩	None	Consumer teaching materials
Forums	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)Forums"⟩	None	Includes mailing list and newsgroup content

permits searchers to access documents that have a keyword as a principal focus. We propose that developers assign as asterisk (*) in front of the subject term that describes the major coverage of the document. The presence of this asterisk will not affect the use of the value by Web crawlers, since they ignore special characters when parsing string values. For example, if "Myocarditis" is the major subject of the document, the previous subject tag would be written as follows:

⟨META NAME = "DC.subject" CONTENT
= "(SCHEME=MCM-MeSHTerm)
*Myocarditis#UI009205"⟩

The MCM-ResourceType Scheme

Users of medical documents need to know not only the subjects covered in the document, as described with MeSH terms above, but also the type of resource. Under the DC framework, resource types are chosen from an enumerated list of types that are optional and repeatable. The highest-level resource types offered by the DC group for Internet-based information sources are text, image, sound, software, data, and interactive. The DC group offers guidelines for defining resource types for specific disciplines (http://sunsite.berkeley.edu/Metadata/types.html). First, resource types should exclude quality parameters, since this involves judgments that should not be encoded in HTML documents themselves. Second, the classification scheme that is used should be intuitive to its audience and comprehensive in scope. Third, the subtyping of resource types should be based on empirical observation of documents that transfer within the discipline.

The Web contains a broader array of medical documents than does traditional medical journal literature. New types of resources include institutional home pages, meeting and conference announcements and schedules, image archives, grant proposals, project descriptions, case reports, and clinicopathologic conferences. In Table 2, we have constrained our resource types to a small core set that can be selected from a pick list. This would facilitate both the assignment of metadata tags and access with a simple search interface. These resource types are presented as suggestions. The breadth and granularity of resource types must ultimately be decided by the community of knowledge users and producers.

The core medical resource types listed in Table 2 have been constructed so that they can be subtyped with more specific resource type descriptions. For example, a pathologist will achieve a more precise search from a heterogeneous database using a resource type such as "pathology images" rather than a more general one such as "images." To support this, it is suggested that in a pathology database indexers subclassify docu-

Table 3 ■

Potential Medical Core Metadata Qualified Resource Types

Qualified Resource Types	MCM Metadata Syntax
Pathology images	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)Image.Path_Image"⟩
Radiographic images	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)Image.Rad_Image"⟩
Clinical images	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)Image.Clinical_Image"⟩
Practice guidelines	⟨META NAME = "DC.type" CONTENT = "(SCHEME=MCM-ResourceType)FAQ.Practice Guidelines"⟩

ments using medical core metadata resource type tags with the following syntax:

⟨META NAME = "DC.type" CONTENT
= "(SCHEME=MCM-ResourceType)
Images.path_images"⟩

The cataloguer could apply content description tags for microscopic images within a pathology database as follows:

⟨META NAME = "DC.type" CONTENT
= (SCHEME=MCM-ResourceType)
Images.path_images.microscopic"⟩

Some additional suggestions for "qualified" resource types are presented in Table 3.

The MCM Relation Element: Item Versus Collection Issues

The Relation element determines the relationship among linked documents. It defines at what level resources should be described by a metadata record and how subdocuments can be made available. For a multiple-page document, such as an image database, it defines how metadata are applied to the Web site's title page and individual images. Our approach for implementing the Relation element is described below. It provides a mechanism for search retrieval algorithms to access parent documents or subdocuments. Our approach specifies that documents from a source have a relationship as *child of* or *parent of* associated documents. By adopting this convention, *child of* documents can be machine processed to inherit descriptive abstracts, publisher, author, and other metadata of parents. Conversely, *parent of* documents can display hyperlinked title metadata of children and subpages within a hierarchic directory structure.

The relation element tag for a diabetic nephrosclerosis pathology image that is one of the images in the University of Utah Webpath Pathology Database would appear as follows:

⟨META=DC.relation Content
= "child_of http://www.webpath.edu/"⟩

An Example of MCM Document Encoding

An example of how this document, if presented to the user as a Web document, might be encoded with MCM is shown below. These metadata tags would be added to the ⟨HEAD⟩ . . . ⟨/HEAD⟩ section of the document as follows:

⟨META NAME = "DC.title" CONTENT = 'Enhancing Internet Medical Document Retrieval with "Medical Core Metadata"'⟩

⟨META NAME = "DC.creator" CONTENT = "Gary Malet, D.O."⟩

⟨META NAME = "DC.creator" CONTENT = "Felix Munoz"⟩

⟨META NAME = "DC.creator" CONTENT = "Richard Appleyard, Ph.D"⟩

⟨META NAME = "DC.creator" CONTENT = "William Hersh, M.D."⟩

⟨META NAME = "DC.subject" CONTENT = "(SCHEME= MCM-MeSHTerm)*Information storage and retrieval #UI016247"⟩

⟨META NAME = "DC.description" CONTENT = "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 . . . . ."⟩

⟨META NAME = "DC.publisher" CONTENT = "Oregon Health Sciences University Division of Medical Informatics & Outcomes Research"⟩

⟨META NAME = "DC.date" CONTENT = "1998-2-5"⟩

⟨META NAME = "DC.type" CONTENT = "Text.Article"⟩

⟨META NAME = "DC.format" CONTENT = "(SCHEME= imt)text/html"⟩

⟨META NAME = "DC.identifier" CONTENT = "http:// medir.ohsu.edu/~metadata/paper.html"⟩

⟨META NAME = "DC.language" CONTENT = "English"⟩

Implementation of Information Retrieval with MCM

The implementation of retrieval systems with MCM will require new tools and software architectures for indexing and retrieval. Some Web-based retrieval tools already collect and manipulate data in a variety of ways, passing it off to other applications in the process. Harvest (http://harvest.transarc.com/afs/transarc.com/public/trg/Harvest/) is an example of a set of tools that allow data to be extracted in customized ways from remote resources to permit construction of subject- or community-specific resources. A number of medical Web crawlers have become available. One of these is Medical World Search, which uses the Harvest program and Perl scripts to allow selection of individual fields from remote Web pages and the presentation of search results in various formats.[6] In its current implementation, it is limited to processing title, URL, and automated abstract metadata. However, it can be programmed with minimal additional effort to process specific metadata field content.

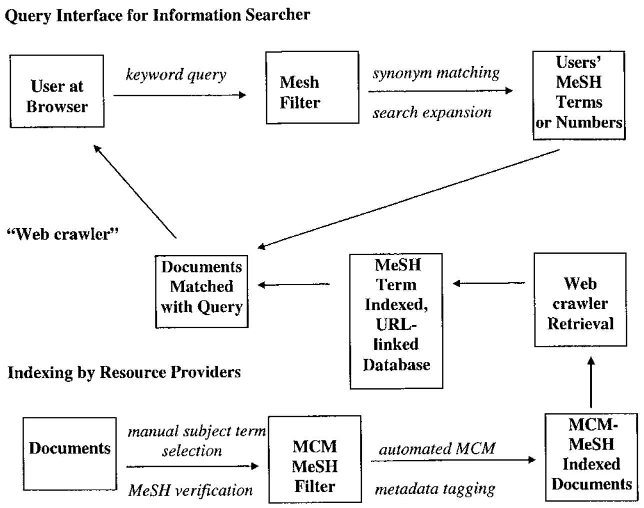
Query Interface for Information Searcher
User at Browser
keyword query
Mesh Filter
synonym matching
search expansion
Users' MeSH Terms or Numbers
"Web crawler"
Documents Matched with Query
MeSH Term Indexed, URL-linked Database
Web crawler Retrieval
Indexing by Resource Providers
Documents
manual subject term selection
MeSH verification
MCM MeSH Filter
automated MCM
metadata tagging
MCM-MeSH Indexed Documents

Figure 1 Architecture for matching Internet subject keyword queries with documents indexed with the medical core metadata syntax.

The MCM-MESH subject metadata tags allow authors to describe the subject of a document to a high level of accuracy and present subject data in a format that both human users and Web crawlers can understand. Tagging this with MeSH numbers metadata syntax will permit a Web crawler to retrieve documents with hierarchic subject content descriptions. The MCM project group has implemented a utility that will interactively verify subject term queries or an indexer's content description against the MeSH controlled language.[7] We expect that MCM-aware search utilities will be able to translate and filter the subject terms that have been assigned. A MEDLINE-type utility could, in turn, expand searches for retrieved documents at related numbers in the hierarchy. Figure 1 represents ways in which information seekers and publishers of medical information might interact.

Future Directions

The metadata schema introduced here is intended to form the foundation of efforts to promote search access to peer-evaluated intellectual efforts distributed on the Internet. The MCM scheme for applying subject content descriptions using MeSH, medical resource types, and relation elements will be submitted to the Dublin Core working group. It is offered as a possible coding system for recognition to standards bodies.

The MCM encodings are presented in the context of Internet resource discovery tools. They can also be applied to classify findings in patients' electronic records. This would allow linkages from patient record to traditional electronic knowledge sources like MEDLINE or could facilitate the use of automated decision tools.

The MCM encodings could allow a search engine to seamlessly connect Internet-based multimedia modules, images, and databases. Indeed, it is conceivable that a global case database could be created that was searchable by disease descriptions and other controlled language tags.

Metadata tags can be utilized by software agents as targets for intelligent processing. They can define the intellectual property rights of Web pages or assist in cataloging hierarchic content relationships for a set of documents. More advanced applications, under consideration by the W^{3}C, which is developing its resource description framework (http://www.w3.org/TR/WD-ref-syntax/), will allow for digital signatures and electronic transactions. Through the adoption of any of these tools, communities of expertise can evolve that will economically control the content of their online information environment. Archives of descriptive metadata for networked resources could be maintained by resource providers or by independent indexing teams. Basic core metadata content descriptions, as outlined in this paper, could be extended by professional indexers. Medical specialties, societies, publishers, and project groups could assume the ownership of metadata encodings and sponsor indexers and search engine developers. In this fashion, they could claim their particular domain of knowledge, idenify quality resources, control the payment for intellectual property, and provide for peer review of their content in a distributed environment. This development would support the publishers and institutions that have maintained the quality of medical research and practice since the beginning of medical science while ensuring that health care providers would be afforded access to the Internet's universal access and multimedia capabilities.

References ■

1. Hersh W. Evidence-based medicine and the Internet. ACP J Club. 1996;5(4):A12–4.
2. Silberg W, Lundberg G, Musacchio R. Assessing, controlling, and assuring the quality of medical information on the Internet: caveat lector et viewor—let the reader and viewer beware. JAMA. 1997;277:1244–5.
3. Brown M. American Interactive Healthcare Professionals Survey. New York: Find/SVP Emerging Technologies Group, 1997.
4. Lowe H, Barnett G. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA. 1994;271:1103–8.
5. Bachrach C, Charen T. Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. Med Inform. 1978;3(3):237–54.
6. Suarez H, Hao X, Chang I. Searching for information on the Internet using the UMLS and Medical World Search. AMIA Annu Fall Symp. 1997:824–8.
7. Munoz F, Hersh W. MCM generator: a JAVA-based tool for generating medical metadata. Proc AMIA Annu Fall Symp. 1998:648–52.