

Winner of LIBRI Best Student Paper Award 1999

The First Monday Metadata Project

ROBIN HENSHAW

Emporia State University, School of Library and Information Management, Emporia, Kansas USA

Publicly available web pages number over 100 million and this number will continue to grow in the foreseeable future. This fact serves to aggravate the increasing difficulty of finding information on the Internet. The application of metatags in the HTML (HyperText Markup Language) header of a web page is one method employed by web site developers and authors to help users locate information on the site. Through the use of a search engine, the user can then locate more relevant matches because metatags relating to the site's contents are used to index and describe the site. A metadata policy for *First Monday*, a peer-reviewed journal on the

Internet, was developed and implemented, with the goal of increasing the ranking of a selected group of articles in some search engines. In the article, metadata and metatags are defined, and an introduction to the Dublin Core Initiative is given. The philosophy and structure of First Monday's metadata is discussed, including syntax-specific examples. The practical application viewpoint taken in this project also necessitated an examination of Internet search engines currently indexing on metatags. The paper concludes with preliminary research conclusions on the utility of the tags to articles in *First Monday*.

Introduction

A recreational user of the Internet may not realise that there are approximately 100 million public web pages (Richmond 1999). The relevance of this figure may not sink in until the user attempts to find a piece of information on the Internet. Knowing the address, or URL, of a Web site, is a decided advantage, but often this information is not readily available. At this point a search engine such as AltaVista, Webcrawler or HotBot must be used. A novice Internet searcher may hold misconceptions about what a search engine does. He may think that every search engine indexes all sites on the World Wide Web, or that typing in several words together will return pages containing all of the requested words. He also expects information to be returned in a clear and indexed fashion (Pollock and Hockley 1997).

Unfortunately, this is not the case. Current search engines index only a fraction of available web sites and use their own set of algorithms to search through those sites (Peterson 1997). To

find more accurate information, a searcher must consider the percentage of the Internet indexed by the search engine of choice, and any special semantics necessary to locate the requested information. "Search Tips" or "Help" pages available on the search engine site can be beneficial in formulating a search strategy, but the results that appear may leave even the most adept online searcher scratching her head.

The problem of locating materials on the World Wide Web has been widely discussed in the journals of many disciplines and in the press, but its resolution is still far from certain (Turner and Brackbill 1998). This paper examines one method that may improve the ranking of a web page in a set of search results – the application of metatags in the HTML header. The project examined selected articles in *First Monday*, a peer-reviewed journal on the Internet (<http://firstmonday.org>). A metadata policy was developed for the publication and metatags were applied to the selected articles in an attempt to raise their ranking when utilising certain Internet search engines.

A definition of metadata and metatags is given, followed by a brief introduction to the Dublin Core Initiative. The philosophy and structure of *First Monday* metadata is discussed with examples of how to apply the syntax to articles, book reviews and the monthly index. A general overview of metatags on the Internet follows including information on those sites currently indexing metadata. The potential utility of the *First Monday* metadata initiative is discussed in the conclusion.

Metadata and metatags

Metadata is the term used to describe data about other data, or attributes of a resource (Turner and Brackbill 1998). Most commonly, it refers to descriptive information about World Wide Web and other networked electronic resources (Dublin Core Metadata Initiative 1998). The need to uncover and exchange electronic information created the need to label or identify every item of information so it is easily retrievable (Paskin 1997). Metadata can be thought of in much the same way as a card catalogue, where each entry describes a resource in the collection. Where as a library's card catalogue is located apart from the actual resource, metadata may also be included in the resource itself. It provides a user with a means to discover that the resource exists, and how it might be obtained or accessed (Turner and Brackbill 1998). Library cataloguing techniques such as the MARC (Machine Readable Catalogue Format) record that work well in a print environment do not always transfer over to their electronic counterparts (Paskin 1997). The metatag is one method of organising metadata that may aid in the resource discovery process. However, there is currently no single agreed way of identifying items of electronic information (Paskin 1997).

Metatags are non-displaying, or hidden, HTML tags that may provide site owners and authors a degree of control over how a web page is indexed. It is intended to be a place to put meta information not defined by other HTML HEAD elements. This allows for better description of the document content for indexing and cataloguing purposes (Turner and Brackbill 1998). Metatags provide a useful way to control web page summaries in some search engines, and can help provide keywords and descriptions on pages that may lack text (Sullivan 1998). It should be re-

membered that currently there is only a framework for metatags in HTML, and nothing that defines exactly what tags exist and how they should be applied. This was one of the challenges in arriving at policies and an applied syntax for *First Monday's* metadata.

An introduction to the Dublin Core Initiative

The Dublin Core Initiative grew out of a metadata workshop sponsored by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) in March, 1995 (Weibel 1995). The group wanted to identify a method of describing networked electronic resources that was more descriptive than an index entry, but less comprehensive than a formal cataloguing record. They were looking for a means to define a "core" set of metadata elements that would allow authors and information providers a way to describe the work and allow for interoperability among "resource discovery tools" such as Internet search engines (Weibel 1995).

Thirteen metadata elements were defined at this first workshop and became known as the Dublin Core Metadata Element Set. Two additional elements were added at a later date, but the basic definitions of the fifteen elements of the Dublin Core (DC) have been unchanged since December, 1996 (Weibel 1999). The elements are: title, creator, subject, description, publisher, other contributors, date, resource type, format, resource identifier, source, language, relation, coverage, rights management. More information on the elements and their descriptions can be found on the Dublin Core Metadata Initiative web site at: http://www.purl.oclc.org/dc/about/element_set.htm.

The Dublin Core (DC) group used seven principles when developing the set of elements. The elements are intrinsic and extensible and also contain the features of syntax independence, optionality (all DC elements are optional), repeatability (any element may be repeated, for example, multiple authors on one work), and modifiability, indicating the element may be redefined to satisfy the needs of individual communities (Weibel 1995). Adhering to these principles can increase the likelihood that the core element set can remain limited and be self-explanatory, but flexible enough to describe a variety of resources (Weibel 1995).

John Kunze recently released an Internet Draft specifying how Dublin Core can be encoded in HTML (Kunze 1999). Kunze states that DC elements are found in the HTML file HEAD, and are expressed using the META and LINK tags. Furthermore, within a META tag the first letter of a Dublin Core element name is capitalised. DC places no restriction on alphabetic case in an element, and any number of META tagged elements may appear together in any order (Kunze 1999). Non-DC elements may also appear together with DC elements (Kunze 1999). A complete copy of his draft is available at: <http://www.ietf.org/internet-drafts/draft-kunze-dchtml-00.txt>.

First Monday's metadata policies

The *First Monday* metadata project looked to develop clear and useful standards for applying metatags to the *First Monday* journal. We wanted the set of elements to be limited, but flexible and easy to modify, particularly since Internet metadata is still in an evolutionary stage. According to Kunze (1999), editing and managing metadata is easier if one style is chosen and followed.

Combining the Dublin Core goals, simplicity of creation and maintenance, and commonly understood semantics, with the practical application of metatags used by today's search engines was another goal of this project. The two most important META tags are the "description" tag and the "keywords" tag in search engines currently supporting metadata (WebPromote 1998). These tags, and others from the Dublin Core element set were combined, arriving at a total of nine elements that could be applied to the monthly index, general articles and book reviews.

How to apply metatags to First Monday articles

The metadata record and the resource it describes may take one of two forms: the elements may be contained in a record separate from the item, for example, a library's card catalogue record, or the metadata may be embedded in the resource itself (Dublin Core Metadata Initiative 1998). The latter method was chosen for use on *First Monday*, meaning each article, book review and monthly index includes its own set of META tags in the HTML header.

As mentioned previously, the metatags assigned to *First Monday* are drawn from the Dublin Core Metadata Element Set. Nine of the fifteen elements are being used on the monthly table of contents, book reviews and general articles. The elements not chosen for inclusion were some still under development or determined not applicable to the journal at this time. Table 1 lists the elements currently applied in the journal, and how to apply the semantics of each.

Explanations for the syntax used with *First Monday* metadata was drawn from research on how search engines use metatags. The metatag named "description" should be kept under 20 words to avoid truncation (MetaTags 1998). This tag gives the search engine the description for the web site. Without this tag, search engines usually take the first text found on the site and make it the description (WebPromote 1998). It is recommended that the "Description" tag be limited to 20–25 words, or about 150–200 characters of text (WebPromote 1998).

Meta "Keywords" tags should avoid excessive repetition, or spamming, as this may result in pages getting lower relevancy in searches, or being banned from the search engine altogether (Meta Tags 1998). The "keywords" metatag tells the search engine exactly under which keywords the site should be searchable. Without this tag, search engines will choose words for the site from the title and text of the site (WebPromote 1998). Keywords should be specific and not overused. Two- and three-word phrases usually work better than single words (WebPromote 1998). As most of the searching population does not capitalise most proper nouns, it is best to keep keywords in lower case (WebPromote 1998). A string of keywords should be separated by a comma followed by a single space (MetaTags 1998).

Another important aspect in applying metadata is to follow a standard format. The three examples below illustrate the syntax used for general articles, book reviews and the monthly index. Noted after the metadata label is a textual description of what should be included for that element and how it should be formatted.

Metatags on the Internet

One goal of applying metatags to the *First Monday* journal is to help the ranking of articles be

Table 1:

Element name	Description	Label	Format
Title	The name given to the resource by the creator or publisher.	TITLE	Sentence case. If a sub-heading is used it should be separated by a colon.
Author	The person or organisation primarily responsible for creating the intellectual content of the resource.	AUTHOR	Title case; "Lastname, Firstname" If multiple authors, names appear in the same order as they are presented on the resource and are separated by a semi-colon and a space.
Description	Textual description of the content of the resource, including abstracts or content descriptions.	DESCRIPTION	Sentence case; Limited to 200 characters, or 20–25 words.
Keywords	The topic of the resource – keywords or phrases describing the subject or content. If a "standardised" keyword is used it should reflect the format noted in the First Monday Keyword List which is being developed.	KEYWORDS	Lower case; proper nouns use title case; Limited to 500 characters, or 50 words or less. Currently, keywords are not assigned in strict order of importance. Multiple keywords are separated by a comma followed by a space.
Other Contributor	A person or organisation not specified in a "creator" element who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organisation specified in a "creator" element (such as an editor, transcriber or illustrator).	CONTRIBUTOR	Title case; "Lastname, Firstname" If multiple contributors, names appear in the same order as they are presented on the resource and are separated by a semi-colon and a space.
Publisher	The entity responsible for making the resource available in its present form.	PUBLISHER	Title case
Date	The date the resource was made available in its present form.	DATE	Follows ISO 8601 Date Profile; "YYYY-MM-DD"
Language	The language in which the resource is available.	LANGUAGE	Lower case; for English, use "en"
Resource Identifier	The URL of the resource	IDENTIFIER	Lower case

Example 1: Article Format

```
<META NAME="Date" CONTENT="yyyy-mm-dd">
<META NAME="Title" CONTENT="The name given to the resource by the creator or publisher">
<META NAME="Author" CONTENT="The person or organisation primarily responsible for creating
the intellectual content of the resource. Format: Lastname, Firstname ; Lastname, Firstname">
<META NAME="Description" CONTENT="Textual description of the content of the resource, including
abstracts or content descriptions. Format: Sentence case; Limited to 200 characters, or 20 – 25 words.">
<META NAME="Keywords" CONTENT="keywords or phrases describing the subject or content.
Format: Lower case; proper nouns use title case; limited to 500 characters, or 50 words or less; multiple
keywords are separated by a comma followed by a space">
<META NAME="Contributor" CONTENT="A person or organisation not specified in a "creator"
element who has made significant intellectual contributions to the resource but whose contribution is
secondary. Format: Title case; "Lastname, Firstname">
<META NAME="Publisher" CONTENT="Name of entity responsible for making the resource
available">
<META NAME="Language" CONTENT="en">
<META NAME="Identifier" CONTENT="(SCHEME=URL) http://enter.url.of.page.here">
```

Example 2: Index Page Format

```
<META NAME="Date" CONTENT="yyyy-mm-dd">
<META NAME="Title" CONTENT="First Monday">
<META NAME="Description" CONTENT="Volume x, Number x. Index to First Monday, peer-reviewed
journal on the Internet">
<META NAME="Keywords" CONTENT="include titles (in lower case) of journal articles for month
being indexed; multiple keywords are separated by a comma followed by a space">
<META NAME="Contributor" CONTENT="all authors and contributors to journal articles for the
month – interviewees.">
<META NAME="Publisher" CONTENT="Name of entity responsible for making the resource
available">
<META NAME="Language" CONTENT="en">
<META NAME="Identifier" CONTENT="(SCHEME=URL) http://enter.url.of.page.here">
```

Notes on the Index Page:

- On the Index page the “Contributor” tag is used instead of the “Author” tag.
- “FM Interviews” are listed under keywords such as “interview” with the person being interviewed noted under the “Contributor” tag.
- “Book Reviews” are listed under a standard *First Monday* keyword, “book review.”

Example 3: Book Review Format

```
<META NAME="Date" CONTENT="yyyy-mm-dd">
<META NAME="Title" CONTENT="First Monday reviews">
<META NAME="Author" CONTENT="This tag should be assigned to those reviewing the books">
<META NAME="Description" CONTENT="Use the subtitle of page or other description as
appropriate.">
<META NAME="Keywords" CONTENT="include titles (in lower case) of resource being reviewed –
separate titles by commas; use standard keyword “book review”; multiple keywords are separated by a
comma followed by a space">
<META NAME="Contributor" CONTENT="Those who authored the resource should be noted with this
tag.">
<META NAME="Publisher" CONTENT="Name of entity responsible for making the resource
available">
<META NAME="Language" CONTENT="en">
<META NAME="Identifier" CONTENT="(SCHEME=URL) http://enter.url.of.page.here">
```

among the top sites in a list of retrieved “hits” when using an Internet search engine. However, not all search engines index metadata. Currently AltaVista, Infoseek and HotBot recognise and use metatags (Turner and Brackbill 1998). A commercial site, MetaTags, indicated that Webcrawler also indexes metatags. A complete chart of search engine features is available on the Search Engine Watch web site at: <http://searchenginewatch.internet.com/webmasters/features.html>. The table is current through December, 1998, and sup-

ports the previous information on which search engines support metatags.

It is important to know what metadata features, if any, are supported in each search engine, to apply the tags more effectively and to improve ranking. AltaVista and Excite’s help screens were consulted to determine their logic for including or excluding metatags. AltaVista indexes both “keyword” and “description” fields, so a search on a combination of words in each will be found (AltaVista 1997). The “description” metatag is re-

turned with the URL as the page summary of a search. They also index up to 1,024 characters of text in both the "description" and "keyword" fields (AltaVista 1997).

Excite's search engine does not index the content of a metatag, but indexes the body text of a web page even if metatags are present (Excite 1998). If a "description" metatag is used Excite will return the content of that tag as the page summary, however, the description must contain valid text about the page it is on (Excite 1998). Excite, Inc. believes this protects their users from unreliable information.

The most important tags for search engine indexing are the "description" and "keywords" tags (Sullivan 1998). In fact, most of the Internet search engines currently supporting metatags recognise only those tags (Turner and Brackbill 1998). A survey on web content was conducted by SiteMetrics Corporation in April, 1998, measuring usage of metatags by 40,000 commercial Web sites owned by 31,000 of the largest U.S. Companies (SiteMetrics Corporation 1998). Overall, only 30% of the sites surveyed use the "keywords" tag and 27% use meta "descriptions". The survey revealed the leaders in assigning "Keyword" metatags were the industrial technology, travel and computer industries, followed by education and utilities. Another interesting find, was that 31% of meta "descriptions" exceeded the recommended 200 character maximum length, and 8% of meta "keywords" exceeded the maximum 1000 character length (SiteMetrics Corporation 1998). These findings were taken into consideration in the application of *First Monday* metatags, where a simple word count was done for the keyword and description elements in each article.

Utility of First Monday's metatags as tools for locating papers in the journal

Research by Turner and Brackbill (1998) found that the use of META tags in the search engines AltaVista and Infoseek does improve rank. Using the "keywords" tag with or without a "description" tag consistently improved a page's retrieval ranking (Turner and Brackbill 1998). However, using only the "description" tag alone did not improve retrieval ranking over Web documents that did not contain any metatags (Turner and

Brackbill 1998). To determine whether the application of metatags affected ranking for articles in the *First Monday*, a preliminary search for approximately fifteen articles was done in the four search engines that support metatags. The ranking and number of hits returned was recorded for those articles. Another search will be done after the tags have been applied and indexed by those search engines in order to determine their effectiveness.

Conclusion

The metatag provides authors and Web site owners a means to control how their information is displayed and retrieved in a search engine. However, it should not be expected that simply adding metatags will necessarily put a page at the top of retrieved sites. To ensure a top ten ranking, special versions of a Web page would need to be designed to suit the idiosyncrasies of each major search engine (Wilson 1998). The metatags applied to *First Monday* follow Dublin Core principals, and were developed with an eye towards flexibility and ease of modification. A heightened interest and general awareness of metadata should alert search engine designers to the need of enable their services to support metatags and other metadata elements currently being discussed by groups such as the Dublin Core Metadata Initiative (Turner and Brackbill 1998).

References

- AltaVista Search Network. 1997. The META tag: Controlling how your Web page is indexed by Alta Vista. URL: http://altavista.digital.com/av/content/addurl_meta.htm.
- Dublin Core Metadata Initiative. 1997. Dublin Core Metadata Element Set: Reference Description. URL: http://purl.oclc.org/dc/about/element_set.htm.
- Dublin Core Metadata Initiative. 1998. A user guide for simple Dublin Core. URL: http://purl.org/DC/documents/working_drafts/wd-guide-current.htm
- Excite Inc. 1998. Understand Meta Tags. URL: <http://www.excite.com/Info/listing8.html>.
- Kunze, J. 1999. Encoding Dublin Core Metadata in HTML. Dublin Core Workshop Series, Internet-draft. URL: <http://www.ietf.org/internet-drafts/draft-kunze-dchtml-00.txt>.
- MetaTags.com. 1998. met.a.tags. URL: <http://www.metatags.com>.

- Paskin, N. 1998. Information Identifiers. Elsevier Science. URL: <http://www.elsevier.co.jp/inca/homepage/about/infoident>.
- Peterson, R. E. 1997. Eight Internet search engines compared. *First Monday* 2(2). URL: http://www.firstmonday.dk/issues/issues2_2/peterson/index.html.
- Pollock, A., A. Hockley. 1997. What's wrong with Internet searching. *D-Lib Magazine* (March). URL: <http://www.dlib.org/dlib/march97/bt/03pollock.html>.
- Richmond, A. 1999. META tagging for search engines. *WDLV.COM*. URL: <http://www.stars.com/Location/Meta/Tag.html>.
- SiteMetrics Corporation. 1998. Web Content Survey. URL: <http://www.sitemetrics.com/contentsurvey>.
- Sullivan, D. 1998. How to use meta tags. Search Engine Watch. URL: <http://searchenginewatch.internet.com/webmasters/meta.html>.
- Sullivan, D. 1998. Search Engine Features Chart. Search Engine Watch. URL: <http://searchenginewatch.internet.com/webmasters/meta.html>.
- Sullivan, D. 1997. The new meta tags are coming – or are they? Search Engine Watch. URL: <http://searchenginewatch.internet.com/sereport/9712-metatags.html>.
- Turner, T. P., and L. Brackbill. 1998. Rising to the top: Evaluating the use of the HTML META tag to improve retrieval of World Wide Web documents through Internet search engines. *Library Resources & Technical Services* 42(4): 258–71.
- WebPromote. 1998. META Tag FAQs. WebPromote. URL: <http://www.webpromote.com/faq/metatagfaqlist.htm>.
- Weibel, S. 1999. The state of the Dublin Core metadata initiative. *D-Lib Magazine* (April). URL: <http://www.dlib.org/dlib/april99/04weibel.html>.
- Weibel, S. 1995. Metadata: The foundations of resource description. *D-Lib Magazine* July. URL: <http://www.dlib.org/dlib/July95/07weibel.html>.
- Wilson, R. F. 1998. How to get higher in the search engines: The science of “Gateway” pages. *Web Marketing Today* (49). URL: <http://www.wilsonweb.com/articles/search-higher.htm>.