# OCLC Investigates Using Classification Tools to Organize Internet Data

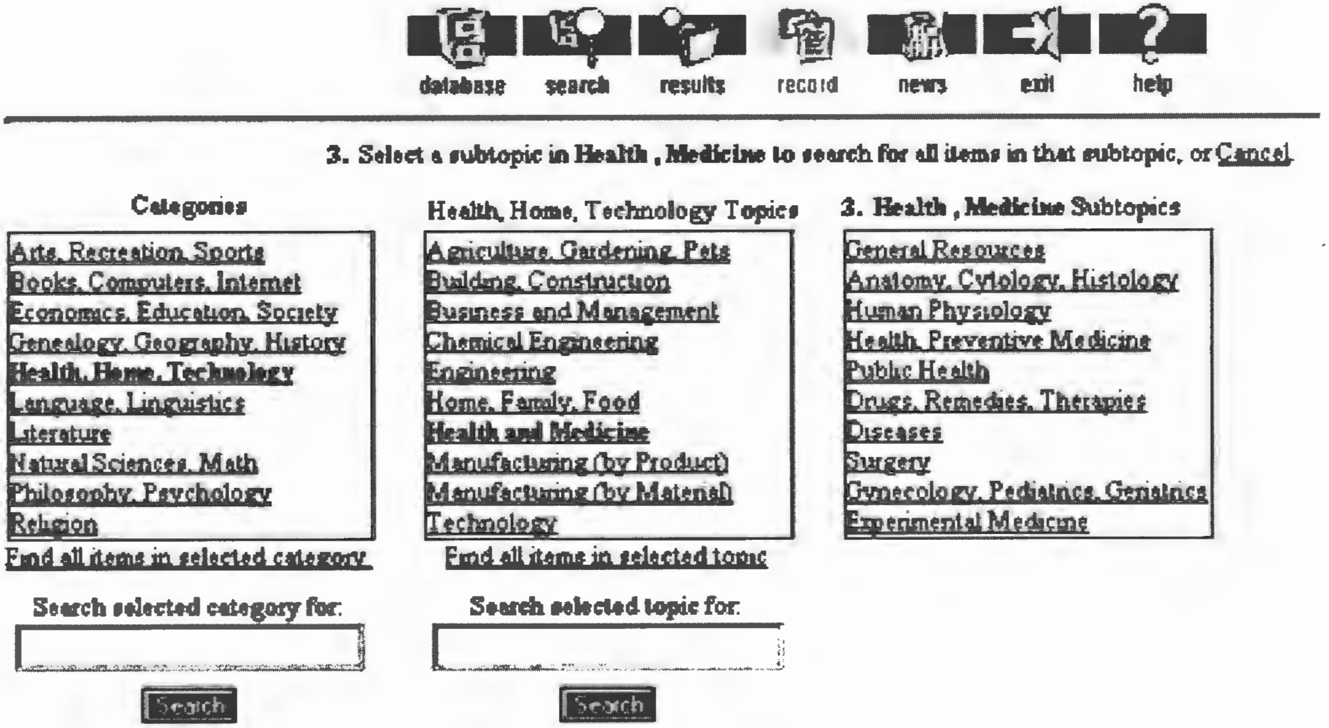
Diane Vizine-Goetz

The knowledge structures that form traditional library classification schemes hold great potential for improving resource description and discovery on the Internet and for organizing electronic document collections. The advantages of assigning subject tokens (classes) to documents from a scheme like the Dewey Decimal Classification (DDC) system are well documented and include:

- providing subject-oriented browsing structures;
- giving context to search terms;
- enabling search refinement;
- providing mechanisms for partitioning and manipulating results sets; and
- enabling multilingual access.

A look at the OCLC NetFirst database will help illustrate some of the advantages of a classified approach to information retrieval. Take, for example, the browsing capability on NetFirst, which provides subject access to Internet-accessible resources using the hierarchical structure of the Dewey Decimal Classification. It allows users to click on subject categories (such as *health, home, technology*), topics (such as *health and medicine*), and subtopics (such as *health, preventive medicine*) to view records grouped by DDC numbers (see Figure 1a).

With just three clicks of the mouse, a set of records numbering nearly 14,000 is reduced to a more manageable set of 249 records (see Figure 1b). Further refinements in searching can be achieved by combining one or more terms with DDC topic categories. For instance, a NetFirst user interested in finding resources containing information about health concerns for travelers can browse to the second level topic *health and medicine* under the category *health, home, technology* and then search for items in



3. Select a subtopic in Health , Medicine to search for all items in that subtopic, or Cancel.

Categories

- Arts, Recreation, Sports
- Books, Computers, Internet
- Economics, Education, Society
- Genealogy, Geography, History
- Health, Home, Technology
- Language, Linguistics
- Literature
- Natural Sciences, Math
- Philosophy, Psychology
- Religion

Find all items in selected category

Search selected category for:

Health, Home, Technology Topics

- Agriculture, Gardening, Pets
- Building, Construction
- Business and Management
- Chemical Engineering
- Engineering
- Home, Family, Food
- Health and Medicine
- Manufacturing (by Product)
- Manufacturing (by Material)
- Technology

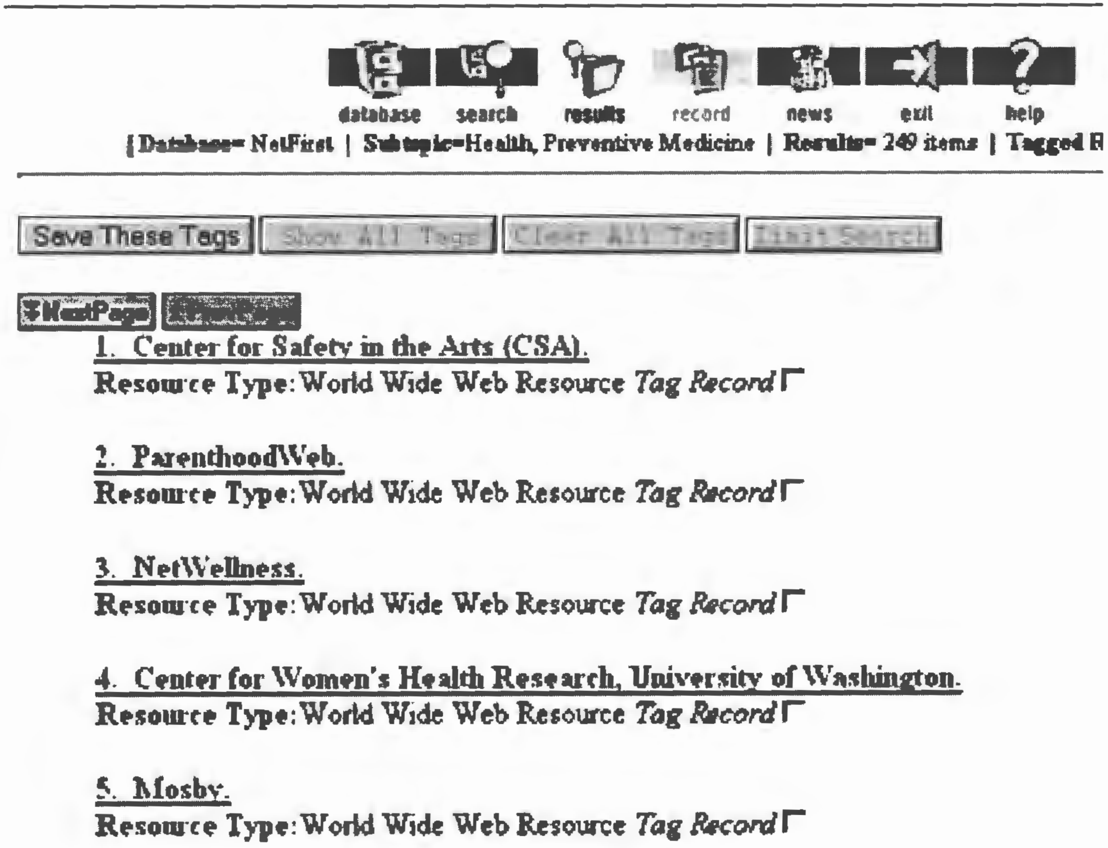
Find all items in selected topic

Search selected topic for:

3. Health , Medicine Subtopics

- General Resources
- Anatomy, Cytology, Histology
- Human Physiology
- Health, Preventive Medicine
- Public Health
- Drugs, Remedies, Therapies
- Diseases
- Surgery
- Gynecology, Pediatrics, Geriatrics
- Experimental Medicine

Figure 1a



|Database= NetFirst | Subtopic=Health, Preventive Medicine | Results= 249 items | Tagged R

1. Center for Safety in the Arts (CSA).
Resource Type: World Wide Web Resource *Tag Record* □

2. ParenthoodWeb.
Resource Type: World Wide Web Resource *Tag Record* □

3. NetWellness.
Resource Type: World Wide Web Resource *Tag Record* □

4. Center for Women's Health Research, University of Washington.
Resource Type: World Wide Web Resource *Tag Record* □

5. Mosby.
Resource Type: World Wide Web Resource *Tag Record* □

Figure 1b

this topic area about travel and tourism (see Figure 2a). Browsing and filtering the database records in this way (using the structure of DDC but not its class numbers) enables users to retrieve relevant items that may not be as easily discovered using traditional keyword searching capabilities. In this case, a keyword search for *health and (travel or tourism)* (see Figure 2b) retrieves 143 items; a similar search filtered by DDC topic area retrieves 25 items, with several potentially relevant items included on the first page of the results display.

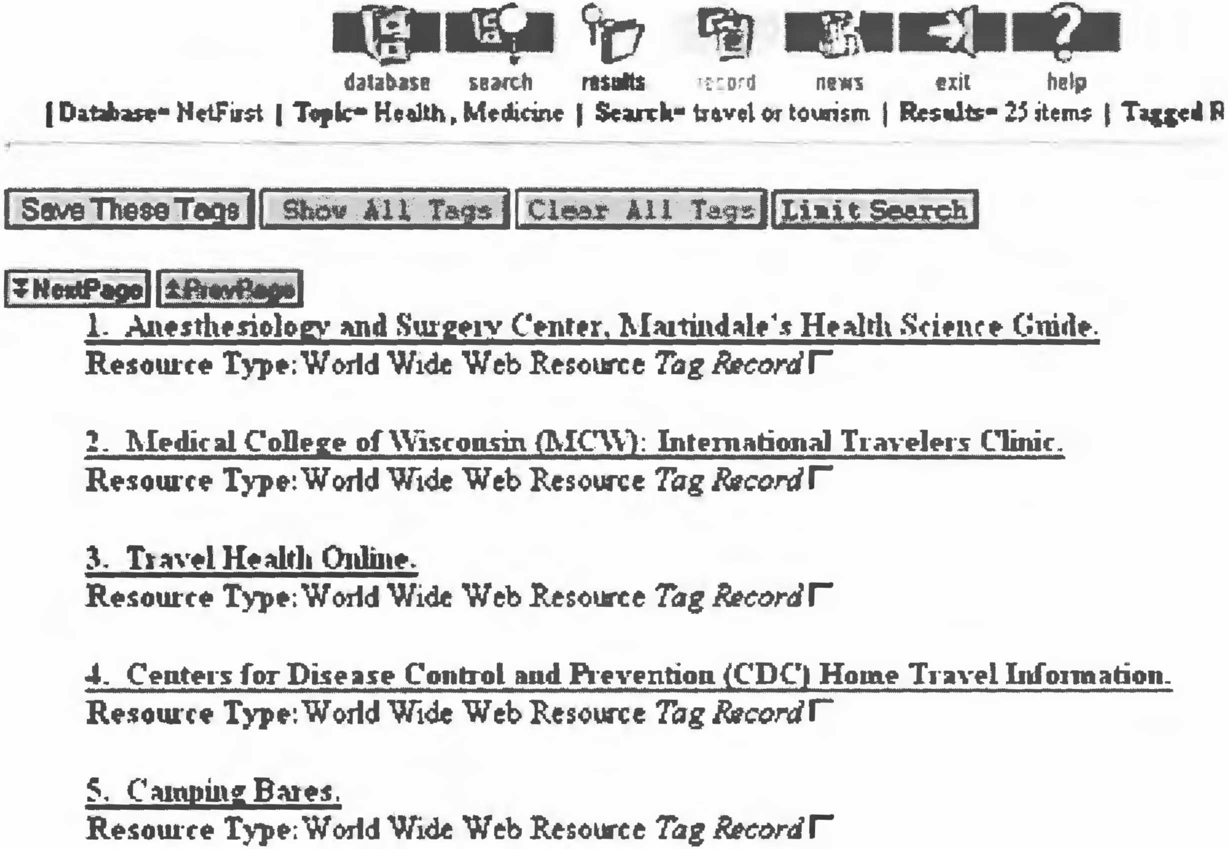
Another example will illustrate some additional benefits of including classification-based subject information in metadata records for electronic documents. Consider the phrase *data mining*, a relatively hot topic that refers to "the process of automatically extracting valid, useful, previously unknown and ultimately comprehensible information from large databases." Although this terminology is not currently used in the Dewey Decimal Classification, the DDC structure can be used to find relevant information. To illustrate, when the keyword search *data mining* or (*data* and *mining*) is run against the NetFirst database, eight items are retrieved on topics ranging from *industrial minerals* and *environmental geotechnology* to *artificial intelligence—databases* and *database management—software*. The titles of the items are:

1. Norsys Software Corporation
2. Ceramic Consulting Group (CCG)
3. Wyoming Technical Information Processing System (WYTIPS), University of Wyoming
4. Colorado School of Mines (CSM)
5. Advanced Visual Systems, SQL
6. d.b. Express
7. Artificial Intelligence Resources
8. Neuralog

The results of this search can be presented to show the broad DDC categories these records fall into, allowing a user to see the various contexts or meanings in which the search terms have been used:

1. Computer software (1 item)
2. Extractive industries (2 items)
3. Geology, hydrology, meteorology (1 item)
4. Information storage and retrieval systems (1 item)
5. Management (1 item)
6. Mining (2 items)

Based on the previous definition of *data mining*, it can be determined that items in the first and fourth categories are potentially relevant. Further



| Database= NetFirst | Topic= Health, Medicine | Search= travel or tourism | Results= 25 items | Tagged R

1. Anesthesiology and Surgery Center, Martindale's Health Science Guide.
Resource Type: World Wide Web Resource *Tag Record* □

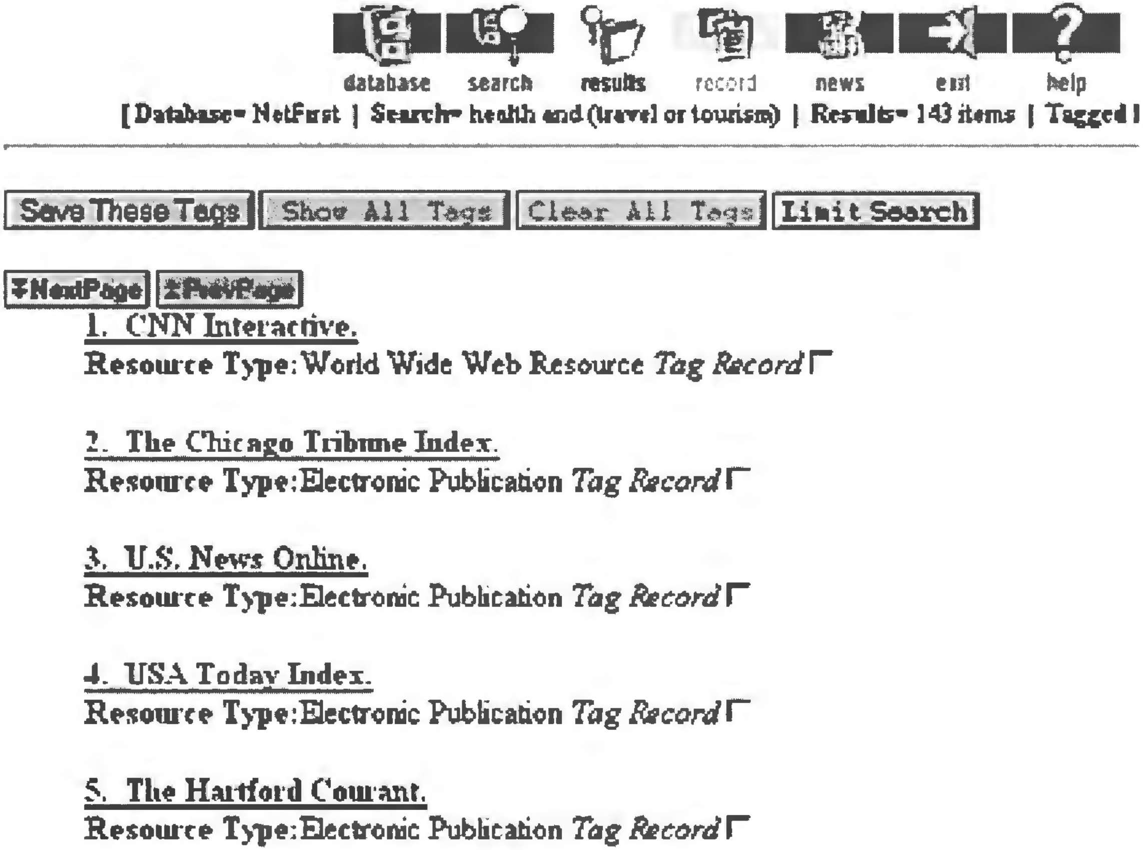
2. Medical College of Wisconsin (MCW): International Travelers Clinic.
Resource Type: World Wide Web Resource *Tag Record* □

3. Travel Health Online.
Resource Type: World Wide Web Resource *Tag Record* □

4. Centers for Disease Control and Prevention (CDC) Home Travel Information.
Resource Type: World Wide Web Resource *Tag Record* □

5. Camping Bares.
Resource Type: World Wide Web Resource *Tag Record* □

Figure 2a



[ Database= NetFirst | Search= health and (travel or tourism) | Results= 143 items | Tagged ]

1. CNN Interactive.
Resource Type: World Wide Web Resource *Tag Record* □

2. The Chicago Tribune Index.
Resource Type: Electronic Publication *Tag Record* □

3. U.S. News Online.
Resource Type: Electronic Publication *Tag Record* □

4. USA Today Index.
Resource Type: Electronic Publication *Tag Record* □

5. The Hartford Courant.
Resource Type: Electronic Publication *Tag Record* □

Figure 2b

search refinements can be enabled by generating information on related topics for DDC classes in relevant records. The NetFirst records for the items in categories one and four contain DDC class numbers 005.3 Computer software, 005.13 Programming languages, 025.06 Information storage and retrieval systems, and 006.3 Artificial intelligence. Using 006.3 as a starting point (see Figure 3), DDC's hierarchical structure can be used to generate coordinate topics and subtopics for use in query reformulation and refinement.

Despite the gains in searching and browsing that can result from using classification data for resource description and discovery, traditional classification schemes are often criticized and then dismissed as Internet organizing tools because of the relatively slow rate new concepts or vocabularies—such as data mining—are assimilated into the systems. Several OCLC-sponsored efforts are underway to improve this situation; two are Office of Research projects—one is ExTended Concept Trees (ETC Trees) and WordSmith and the other is an ongoing service of OCLC Forest Press. In the latter, the Dewey editorial staff review newly approved Library of Congress Subject Headings (LCSH) and pair these with candidate DDC numbers. These new headings represent topics of current interest not specifically mentioned in the latest edition of the DDC. The WordSmith project involves building a set of natural language parsing tools for use in OCLC research projects. WordSmith tools are being used to enhance the DDC with supplemental vocabulary from free text.

ETC Trees is the major project devoted to expanding the Dewey knowledge base. The goal of the project is to augment Dewey concept trees with supplemental vocabulary and to extend these structures through associations with other subject-oriented knowledge bases. Linking the DDC with other subject-access systems can provide:

- useful index terms not found in terminology used in Dewey;
- a mechanism for associating new topics with the classification; and
- navigation and retrieval tools based on outlines of knowledge of other systems.

The imported terminology and other associations are then combined with the Dewey knowledge base to automatically assign subjects to electronic documents. An example of a Dewey extended concept tree is shown in Figure 4.

ExTended Concept Trees is largely directed toward exploiting technology to link subject-access systems like LCSH and the Library of Congress Classification with the DDC. Linking is accomplished by mining WorldCat (the OCLC Online Union Catalog) and electronic versions of other subject access systems for relationships between subject-oriented data in these files and the Dewey knowledge structure. The techniques for making these associations include use of OCLC's Scorpion system.

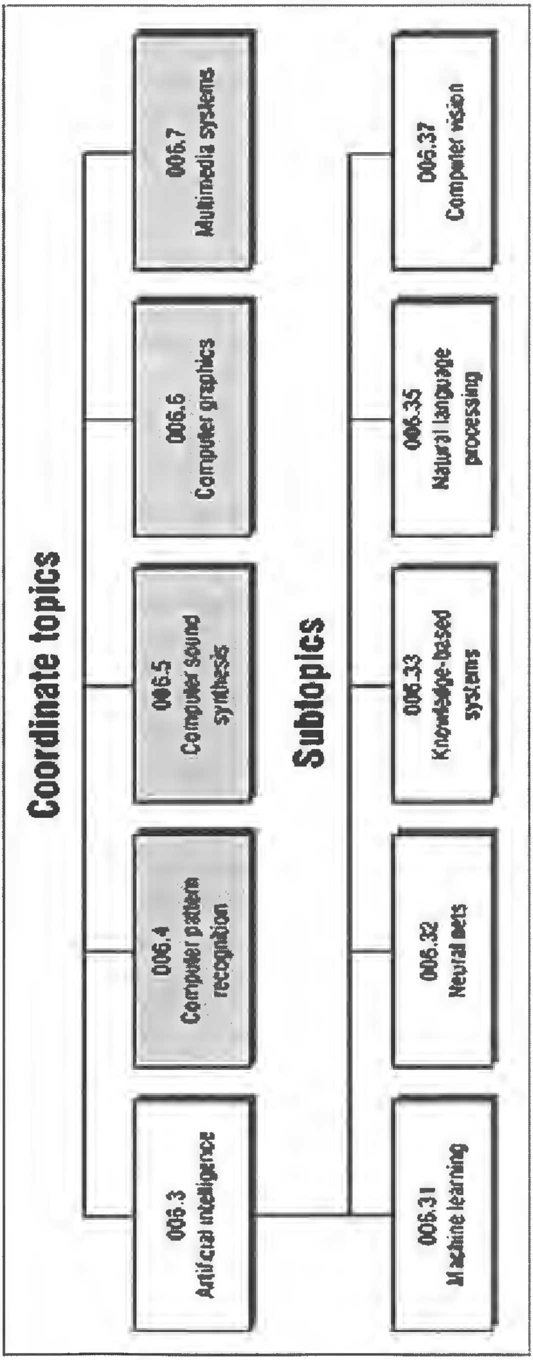

Coordinate topics
006.3
Artificial intelligence
006.4
Computer pattern recognition
006.5
Computer sound synthesis
006.6
Computer graphics
006.7
Multimedia systems
Subtopics
006.31
Machine learning
006.32
Neural nets
006.33
Knowledge-based systems
006.35
Natural language processing
006.37
Computer vision

Figure 3

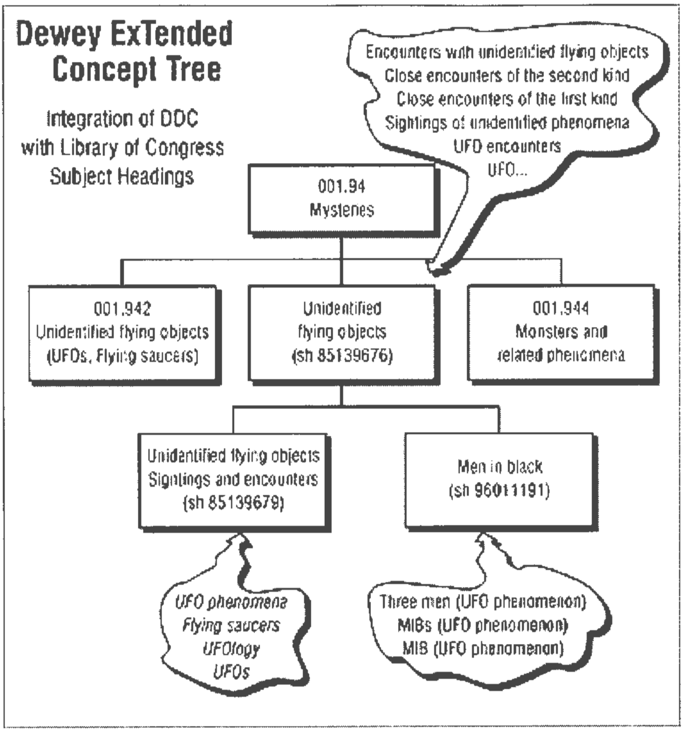
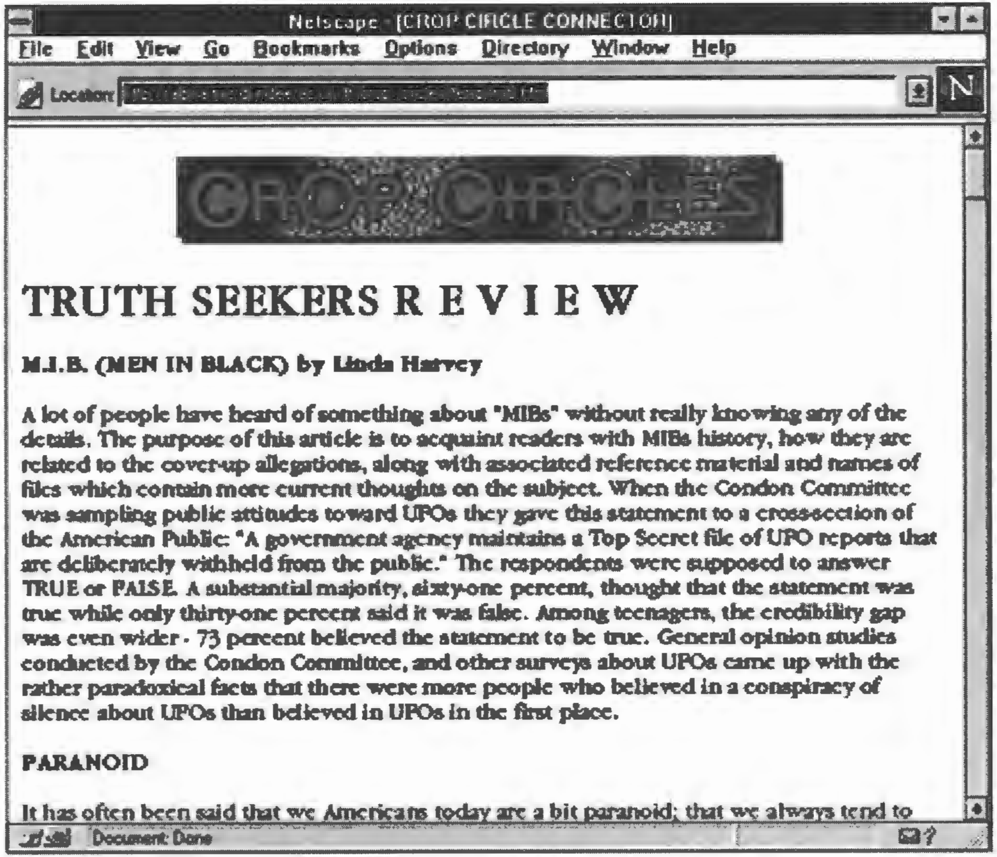

Dewey ExTended Concept Tree
Integration of DDC with Library of Congress Subject Headings
Encounters with unidentified flying objects
Close encounters of the second kind
Close encounters of the first kind
Sightings of unidentified phenomena
UFO encounters
UFO...
001.94 Mysteries
001.942 Unidentified flying objects (UFOs, Flying saucers)
Unidentified flying objects (sh 85139676)
001.944 Monsters and related phenomena
Unidentified flying objects Sightings and encounters (sh 85139679)
Men in black (sh 96011191)
UFO phenomena
Flying saucers
UFOlogy
UFOs
Three men (UFO phenomenon)
MIBs (UFO phenomenon)
MIB (UFO phenomenon)

Figure 4

Scorpion is a research prototype that employs a series of ranked retrieval databases built from the machine-readable version of DDC 21. The system generates ranked lists of Dewey numbers that function as possible subject descriptors for documents. The Scorpion databases can be accessed via a Web interface that is capable of retrieving an electronic document and generating a database query from its content. For example, when a Web document, in this case *M.I.B. (MEN IN BLACK)* by Linda Harvey, is processed by the Scorpion system, results like those shown in Figure 5 are produced. The highest ranked class assigned to this document is 001.94 Mysteries (see Figure 6a). The Scorpion system record for this class number is shown in Figure 6b. The highlighted terms indicate matches between terminology in the input document and in the Scorpion classification records. Observe the "class here note" at the end of the record that instructs DDC users to apply this class number to items about *nonastronomical extraterrestrial influences on earth.* The two related class numbers—001.942 Unidentified flying objects (UFOs, Flying saucers) and 001.944 Monsters and related phenomena—are also among the top twenty classes assigned by the system. This example illustrates the potential value of the Scorpion system to automatically generate subject-oriented metadata for electronic documents.



# TRUTH SEEKERS REVIEW

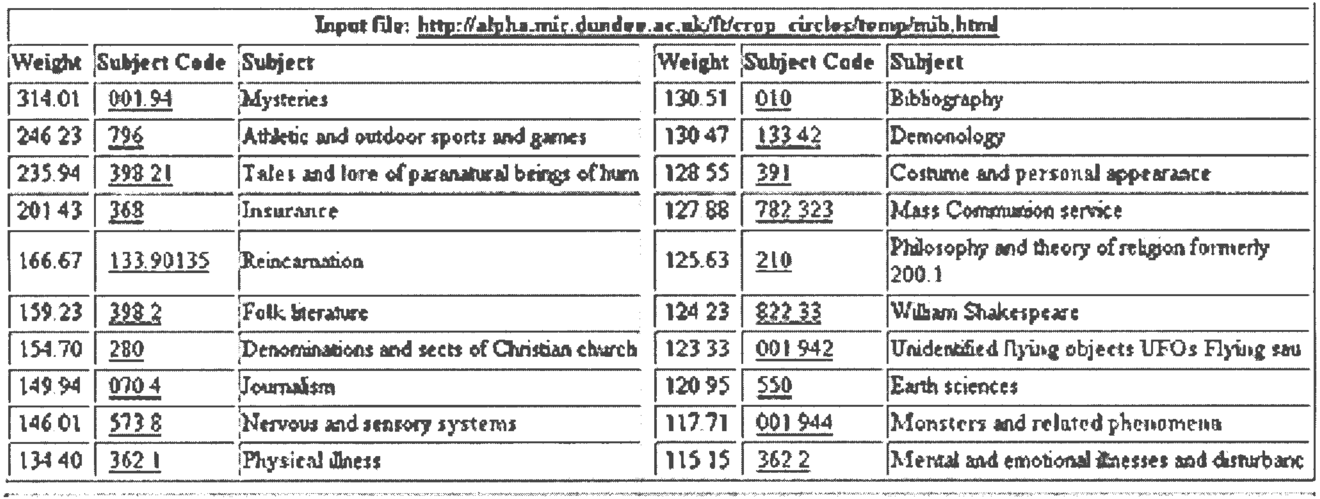
**M.I.B. (MEN IN BLACK) by Linda Harvey**

A lot of people have heard of something about "MIBs" without really knowing any of the details. The purpose of this article is to acquaint readers with MIBs history, how they are related to the cover-up allegations, along with associated reference material and names of files which contain more current thoughts on the subject. When the Condon Committee was sampling public attitudes toward UFOs they gave this statement to a crosssection of the American Public: "A government agency maintains a Top Secret file of UFO reports that are deliberately withheld from the public." The respondents were supposed to answer TRUE or FALSE. A substantial majority, sixty-one percent, thought that the statement was true while only thirty-one percent said it was false. Among teenagers, the credibility gap was even wider - 73 percent believed the statement to be true. General opinion studies conducted by the Condon Committee, and other surveys about UFOs came up with the rather paradoxical facts that there were more people who believed in a conspiracy of silence about UFOs than believed in UFOs in the first place.

**PARANOID**

It has often been said that we Americans today are a bit paranoid; that we always tend to

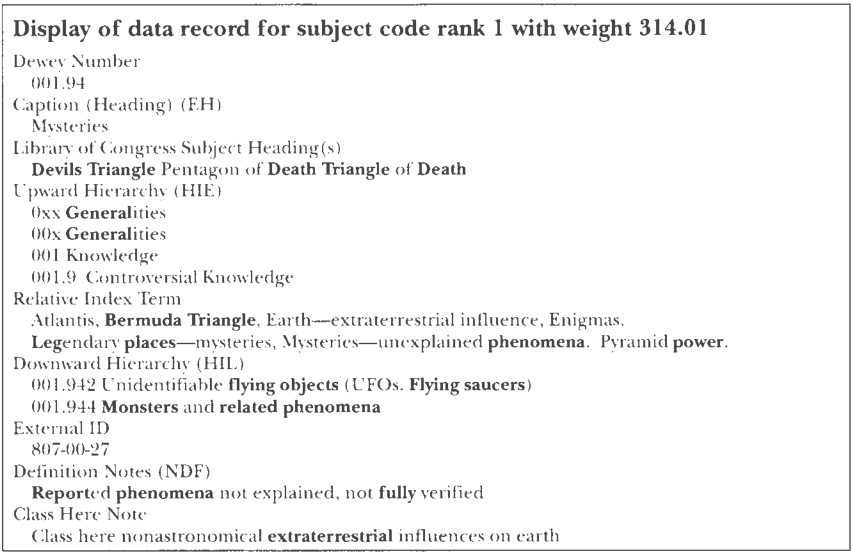
Figure 5



Input file: http://alpha.mic.dundee.ac.uk/ft/crop_circles/temp/mib.html

| Weight | Subject Code | Subject | Weight | Subject Code | Subject |
|---|---|---|---|---|---|
| 314.01 | 001.94 | Mysteries | 130.51 | 010 | Bibliography |
| 246 23 | 796 | Athletic and outdoor sports and games | 130 47 | 133 42 | Demonology |
| 235.94 | 398 21 | Tales and lore of paranatural beings of hum | 128 55 | 391 | Costume and personal appearance |
| 201 43 | 368 | Insurance | 127 88 | 782 323 | Mass Communion service |
| 166.67 | 133.90135 | Reincarnation | 125.63 | 210 | Philosophy and theory of religion formerly 200.1 |
| 159.23 | 398 2 | Folk literature | 124 23 | 822 33 | William Shakespeare |
| 154.70 | 280 | Denominations and sects of Christian church | 123 33 | 001 942 | Unidentified flying objects UFOs Flying sau |
| 149 94 | 070 4 | Journalism | 120 95 | 550 | Earth sciences |
| 146 01 | 573 8 | Nervous and sensory systems | 117.71 | 001 944 | Monsters and related phenomena |
| 134 40 | 362 1 | Physical illness | 115 15 | 362 2 | Mental and emotional illnesses and disturbanc |

Figure 6a



**Display of data record for subject code rank 1 with weight 314.01**

Dewey Number

001.94

Caption (Heading) (EH)

Mysteries

Library of Congress Subject Heading(s)

**Devils Triangle** Pentagon of **Death Triangle** of **Death**

Upward Hierarchy (HIE)

0xx **Generali**ties

00x **Generali**ties

001 Knowledge

001.9 Controversial Knowledge

Relative Index Term

Atlantis, **Bermuda Triangle**, Earth—extraterrestrial influence, Enigmas, **Legend**ary **places**—mysteries, Mysteries—unexplained **phenomena**. Pyramid **power**.

Downward Hierarchy (HIL)

001.942 Unidentifiable **flying objects** (UFOs. **Flying saucers**)

001.944 **Monsters** and **related phenomena**

External ID

807-00-27

Definition Notes (NDF)

**Report**ed **phenomena** not explained, not **fully** verified

Class Here Note

Class here nonastronomical **extraterrestrial** influences on earth

Figure 6b

Staying with the topic *Men in Black*, one additional example shows a technique being explored to affect automatic associations between the DDC knowledge base and other subject access systems. Since this topic corresponds to the LC subject heading *Men in Black (UFO phenomenon)*, it is possible to generate a "concept record" for the topic from information in the OCLC Authority File (see Figure 7).

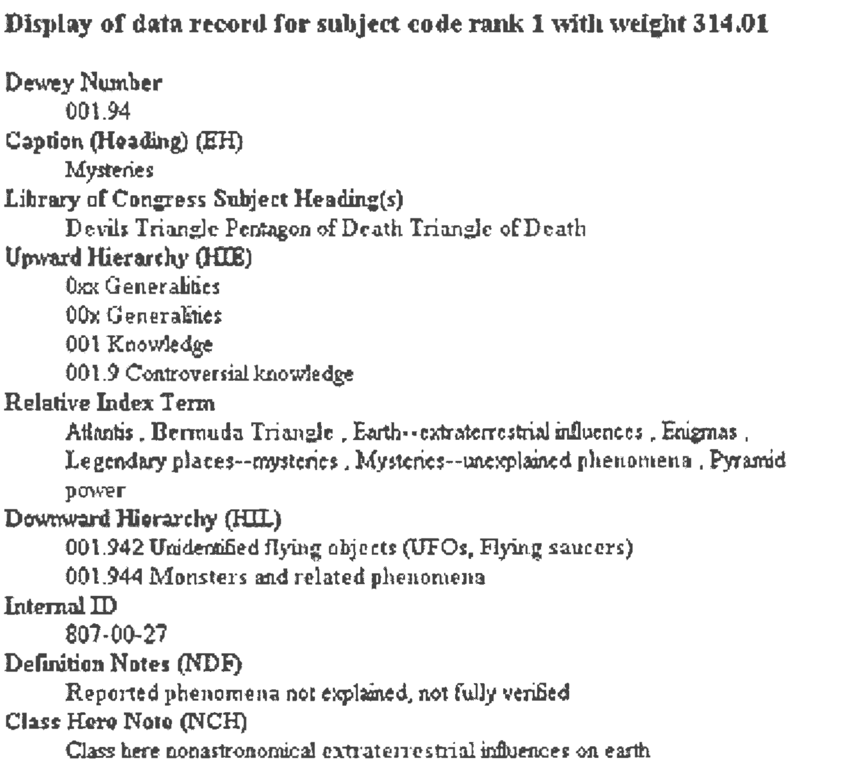
An HTML version of the concept record is generated and then sent in turn to the Scorpion system for processing, with the following top three classifications being returned:

| **Dewey Number** | **Caption (Heading)** |
|---|---|
| 001.942 | Unidentified flying objects (UFOs, Flying saucers) |
| 133.88 | Psychokinesis |
| 001.94 | Mysteries |

These and similar results are quite promising (the candidate DDC class paired with this heading by the Dewey editors is 001.942), but many research questions remain:

- How should information from discrete knowledge bases be integrated?
- What are the relationships among mapped concepts and how should they be coded?
- How can Scorpion results sets be post-processed to filter out spurious classes and collocate valid ones?

In spite of these challenges, it is important to pursue research into automatic assignment of subjects from classification-grounded knowledge bases, since this approach may play a critical role in providing conceptual structuring for large collections of electronic documents with mutable content. By including classification-based subject tokens in metadata records, many advanced browsing and retrieval capabilities can also be provided.



**Display of data record for subject code rank 1 with weight 314.01**

**Dewey Number**
001.94
**Caption (Heading) (EH)**
Mysteries
**Library of Congress Subject Heading(s)**
Devils Triangle Pentagon of Death Triangle of Death
**Upward Hierarchy (HIE)**
0xx Generalities
00x Generalities
001 Knowledge
001.9 Controversial knowledge
**Relative Index Term**
Atlantis , Bermuda Triangle , Earth--extraterrestrial influences , Enigmas , Legendary places--mysteries , Mysteries--unexplained phenomena , Pyramid power
**Downward Hierarchy (HIL)**
001.942 Unidentified flying objects (UFOs, Flying saucers)
001.944 Monsters and related phenomena
**Internal ID**
807-00-27
**Definition Notes (NDF)**
Reported phenomena not explained, not fully verified
**Class Here Note (NCH)**
Class here nonastronomical extraterrestrial influences on earth

Figure 7