Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

Use and Management of Classification Systems for Knowledge-Based Indexing

Susanne M. Humphrey
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, MD 20894

The MedIndEx (Medical Indexing Expert) research project combines artificial intelligence and information retrieval principles and methods to develop and test an interactive knowledge-based prototype for computer-assisted indexing of the MEDLINE® database. By encoding the indexing scheme in a knowledge base (KB), and designing a system for indexers to use in a workstation environment, the objective of this project is to facilitate "expert indexing" that is performed at the National Library of Medicine

1. INTRODUCTION

MedIndEx is a prototype knowledge-based expert system designed to assist indexers in creating subject access points for performing MeSH thesaurus-based searches of the MEDLINE database. MEDLINE contains more than 6.6 million citations covering the periodical biomedical literature since 1966. According to its fiscal year 1990 report [1], NLM indexed nearly 400,000 documents for MEDLINE. Over two million online searches were performed on the current MEDLINE database (the most recent 2-3 years) using NLM's retrieval system. This does not include searches of MEDLINE on commercial retrieval systems that lease the database.

The main objective of MedIndEx is to develop interactive knowledge-based systems to facilitate expert indexing that goes into the MEDLINE product. The system consists of computer representations of indexing concepts and executable rules. It is designed for a sophisticated workstation environment. A second objective, that has evolved from this project, is to utilize the same representations and environment to develop intelligent retrieval systems.

A 250-page technical report, which includes the system design document and indexing manual, is available from NTIS [2].

2. CONVENTIONAL INDEXING

Before describing MedIndEx, we will summarize characteristics of conventional indexing that would suggest a knowledge-based approach as a natural outgrowth.

MeSH contains about 16,000 headings, 80 subheadings which form heading-subheading pre-coordinations as indexing terms, and about 60,000 supplementary chemical terms which serve to enlarge the entry vocabulary to regular MeSH chemical headings. In addition, MeSH contains official alternate terms, like synonyms, abbreviations, and other variants. Although MeSH is in machine-readable form, indexers use published forms to do their work. Tools that contain indexing rules *per se* include the alphabetic MeSH, which has annotations appended to individual terms, the indexing manual, technical notes, and other publications.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

Data entry for indexing is interactive. Indexers enter indexing terms at computer terminals linked to an IBM mainframe. The computer does do a few very important things. It validates terms that are entered. It substitutes entries with the preferred form. It validates heading-subheading coordinations using general rules of permissible combinations. And it issues simple warnings, for instance, having to do with checktags. For example, when an indexer enters the term Adult, the system will provide the term Human. But for really knowledge-intensive assistance, indexers rely on publications and their training.

Fundamental indexing tenets include specificity and multiplicity. These correspond to the following fundamental rules of indexing identified by Lancaster [3]:

- Include all the topics known to be of interest to the users of the information service that are treated substantively in the document.

- Index each of these as specifically as the vocabulary of the system allows and the needs or interests of the users warrant.

The first rule corresponds to the conceptual analysis stage of indexing; the second, to the translation stage [4].

MeSH is unique among conventional thesauri of its size in having a unifying classification scheme, i.e., the MeSH trees. The trees are used for finding the most specific term for a topic. Specificity is also achieved by use of single-term pre-coordinates. In many cases, these demonstrate inherent relations that might be quite useful in an expert system. An example would be BODY-SITE, inherent in the headings Lung Diseases (= Disease BODY-SITE Lung), Angiography (= Radiography BODY-SITE Arteries), and numerous others. The classification is also used for rules of coordinate indexing, whereby a concept is indexed by assigning one or more descriptors, each coming from a different node in the classification. Rules for permissible heading-subheading pre-coordinations again follow the MeSH classification scheme. For example, the subheading "cytology" may qualify only those terms in the anatomy and organism category. We furthermore note that the indexing manual itself is largely organized by major MeSH tree categories (Anatomy, Organisms, Diseases, Chemicals and Drugs, and so forth).

Coordinate indexing also achieves multiplicity. An example taken from the indexing manual is the topic *anticonvulsant therapy of epilepsy causing abnormalities*, indexed by Anticonvulsants/ ADVERSE EFFECTS + Abnormalities, Drug-Induced + Epilepsy/DRUG THERAPY. Note this includes pre-coordinations (single-term and mainheading-subheading) as well as coordination of indexing terms. This example suggests several relations that might be useful for assistance in achieving these sorts of coordination, such as ADVERSE-EFFECT (Anticonvulsants ADVERSE-EFFECT Abnormalities, Drug-Induced), ETIOLOGY (Abnormalities, Drug-Induced ETIOLOGY Anticonvulsants), and so forth.

Finally, conventional thesauri make the following distinction that carries over to knowledge-based systems: some sorts of assistance are to be *prescriptive*, the remaining merely *suggestive* [5].

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

To illustrate conventional indexing, and then later MedIndEx, we will use a real-world MEDLINE search query on the subject of estrogen replacement therapy in relation to osteoporosis or heart disease in postmenopausal women, focusing on the following citation not retrieved by a search strategy for this query because of omission of any Osteoporosis indexing term:

TI - Monitoring skeletal response to estrogen.
AB - Estrogen replacement therapy at accepted doses is not fully effective in preventing bone loss and fractures in postmenopausal women. Bone densitometry is useful for monitoring estrogen replacement therapy to assess dose, foster compliance, and check for secondary bone loss. The most appropriate site for bone loss monitoring is probably the spine because it shows larger decreases at the menopause than appendicular sites, it shows larger increases with therapy, and it has clinical import in terms of fracture. Both dual-photon absorptiometry (or dual-energy x-ray absorptiometry) and computed tomography are the preferred monitoring methods. The precision of these densitometry methods is generally adequate to permit interim decisions with regard to continuing therapy, as well as conclusive decisions on therapeutic efficacy after 1 to 2 years of monitoring. Judicious use of densitometry in combination with biochemical determinations can enhance therapeutic control and provide both patient and physician confidence in long-term estrogen replacement therapy.
MH - Absorptiometry, Photon; Bone Density/*DRUG EFFECTS; *Estrogen Replacement Therapy; Female; Human; Menopause/PHYSIOLOGY; Monitoring, Physiologic; Tomography, X-Ray Computed

The published MeSH has two specific aids that might have lead to use of Osteoporosis, Postmenopausal for indexing this document. In particular, *bone loss in postmenopausal women* (see the first sentence in the abstract) is comparable to the MeSH cross-reference "Bone Loss, Postmenopausal SEE Osteoporosis, Postmenopausal" (SEE cross-references in MeSH are presumed to be prescriptive). (It should be emphasized that this phrase mentioned only in the abstract would not be sufficient grounds for picking up this concept for indexing. In order to be indexable, concepts in abstracts must be substantively discussed in the actual text. In fact, indexers are instructed not to read the abstract until they are finished indexing the article, and to do this only as a check that they have not missed anything discussed in the text. In this case, this concept was also found in the text.) Furthermore, MeSH refers to the omitted term in the cross-reference "Menopause SEE RELATED Osteoporosis, Postmenopausal". That these published indexing aids might well not have been consulted is not surprising, as there is little time during indexing for looking up all text phrases in a document, or for looking up terms, like Menopause, that are known to be in MeSH simply in order to see what they refer to as related terms. (The document contained additional clues but these were in peripheral sections: "osteoporosis" was among the three author-supplied key words, others being "bone densitometry" and "estrogen", and in one-fifth of titles in the sixty end-references, including the first three).

This example was intended to highlight the difficulties in using a system where the computer does not provide much in the way of interactive assistance, but also to show that the conventional system does provide a foundation to build on.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

3. MedIndEx INDEXING

We turn now to use of MedIndEx and how it addresses these indexing problems. But first, it seems useful to emphasize what MedIndEx is *not*. It is not a new set of concepts for biomedicine (provided by MeSH), nor is it a new indexing scheme (already in place at NLM). Although an output of the system is an indexing database that, unlike conventional indexing, contains explicit linkages between concepts, these relationships, encoded as *indexing frames*, are required in order for the system to give situation-specific assistance to indexers as they interact with the system. These frames also provide information for the system to automatically generate conventional indexing. It is this second output that would be used to evaluate the prototype, and which would be of immediate practical use in conventional MEDLINE indexing if the system were adopted. (Of course, the system design does not constrain developers from establishing whatever linkages they wish.)

In contrast to conventional indexing, MedIndEx provides indexing rules as executable computer code, explicit domain-specific relations subdividing concepts, and organization of concepts in an *inheritance classification*. These are not provided by the conventional system, and efforts in connection with them are part of our research. Specifically, the MeSH classification is not suitable for supporting inheritance; MeSH does not contain specific, explicit relations for linking concepts; and MeSH and other indexing tools give advice to users in the form of text to be looked up and read, rather than encoding it as procedural knowledge to assist users interactively. Thus, in order to go from a thesaurus-based to a knowledge-based system, we need to develop existing classifications and suggested, but inexplicit, relations much further, for a more rigorous form of computational use.

MedIndEx uses a Lisp-based frame language and data structure. (A brief tutorial using ten sample frames, and referencing some sources on frames, appears in [6].) The use of frames is an object-oriented approach that supports expression of relationships between concepts, encoding and application of procedural knowledge, inheritance of procedures and data, and internal retrieval for accessing and displaying data from other locations (in the same frame or other frames). We have developed an experimental frame language, giving us the much-needed flexibility of a research environment.

Knowledge-base frames are subdivided by slots (relations), and slots are subdivided by *facets* which encode various types of knowledge-based assistance, such as that needed to fill slots. Each indexing frame is an instance of a knowledge-base (KB) frame, and is named by the concatenation of this KB frame-term with a unique document identifier for the document it indexes. Being an instance of a KB frame means that the indexing frame inherits procedures and possibly data from this KB frame. Essentially, all system actions that occur while indexers fill an indexing frame are encoded in the knowledge base. Some actions are global, i.e., apply to filling of any frame, and are encoded or called by procedures in a frame named KB-ROOT, which is more like a system frame than a KB frame. Other, domain-dependent actions are encoded or called from frames with names for diseases, procedures, substances, organisms, and so forth.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

This frame representation, together with the interface, guides indexers in filling indexing frames which are instances of knowledge base (KB) frames. Based on these indexing frames, the system automatically generates conventional indexing output. This output is important for two reasons: It is a form that is directly usable as access points to the conventional MEDLINE database, and it can be used for evaluating the system using algorithms that compare versions of conventional indexing.

Following the example begun in the previous section, ideally citations relevant to the sample query should be indexed to the term Estrogen Replacement Therapy. If this is the initial term selected by an indexer using MedIndEx, the system will display an Estrogen Replacement Therapy indexing frame for the indexer to fill. (The system does not bother displaying the document number portion of the name of indexing frames.) As shown in Figure 1, the indexer has entered the filler Osteoporosis, Postmenopausal in the first slot, PROBLEM, which is displayed in the window labeled Current Slot. The Current Frame window contains the remaining slots, and marks the location of the slot that is currently being filled. This use of slots as prompts for indexing terms is considered a form of assistance to help ensure more complete indexing, as it focuses indexers' attention on aspects of topics that should be considered.

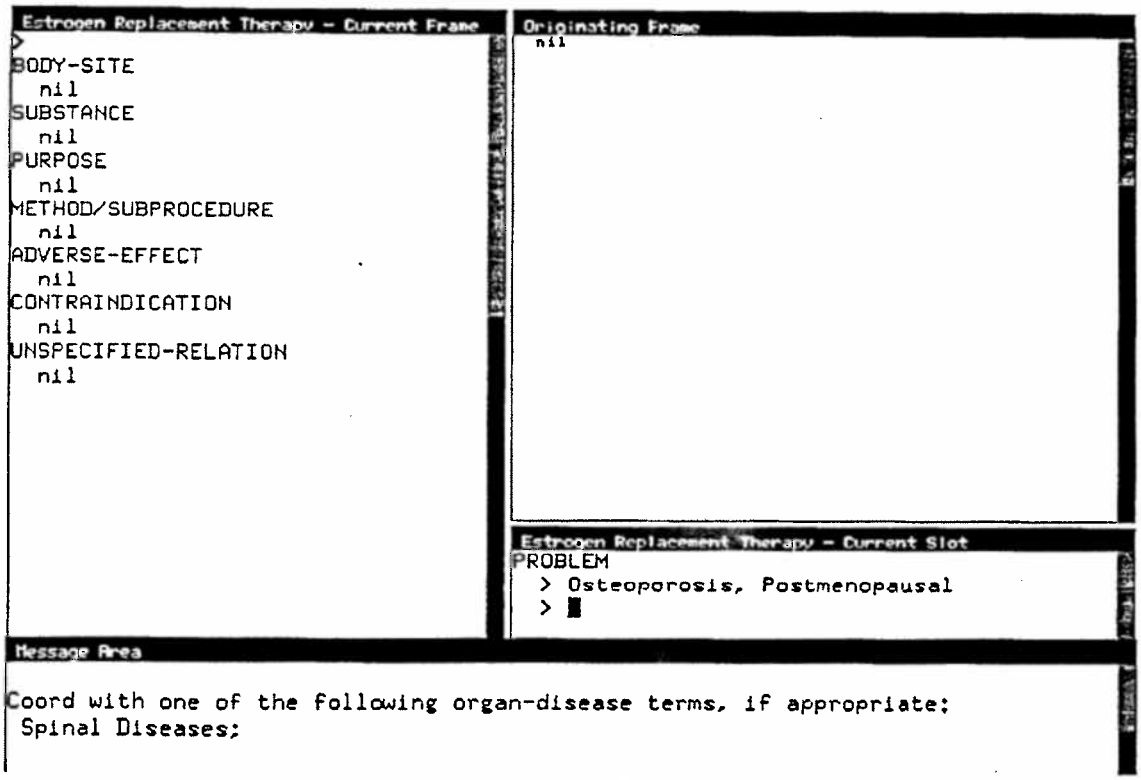

Figure 1. Indexing frame showing *coord* message subsequent to filling current slot.

Another form of assistance is display of system messages in the Message Area window. In this figure the message reminds the indexer to consider adding another filler to the current slot, namely,

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

the organ-disease term Spinal Diseases. These sorts of *coord* messages are patterned after similar MeSH annotations (notes to indexers, catalogers, and searchers, as part of individual entries in *Alphabetic MeSH*). The advantage of a frame-based computer system is that procedures may be encoded in a high-level frame, and executed via accessing this code through inheritance links in the KB. For instance, the code that resulted in the current message appears in the PROBLEM slot of the Procedures frame, and is accessed by the current indexing frame via inheritance, first from the Estrogen Replacement Therapy KB frame, and then inheritance links from this KB frame ultimately to the Procedures frame. Arguments for the message, e.g., the suggested term Spinal Diseases, may be stored as data in KB frames. This message, furthermore, falls into the class of suggestive assistance, i.e., the indexer may ignore the suggestion.

Figure 2 shows the PROBLEM slot and filler returned to the Current Frame window, with the next slot prompt for SUBSTANCE in relation to the previously-entered problem Osteoporosis, Postmenopausal. The filler Estrogens is displayed by the system automatically as a default. This sort of assistance, where the system displays so-called *candidate* fillers (not necessarily KB defaults, but may be based on previously-entered data in indexing frames for the current document), is another form of suggestive assistance, e.g., the indexer may replace this filler with a specific estrogen term, like Estradiol.

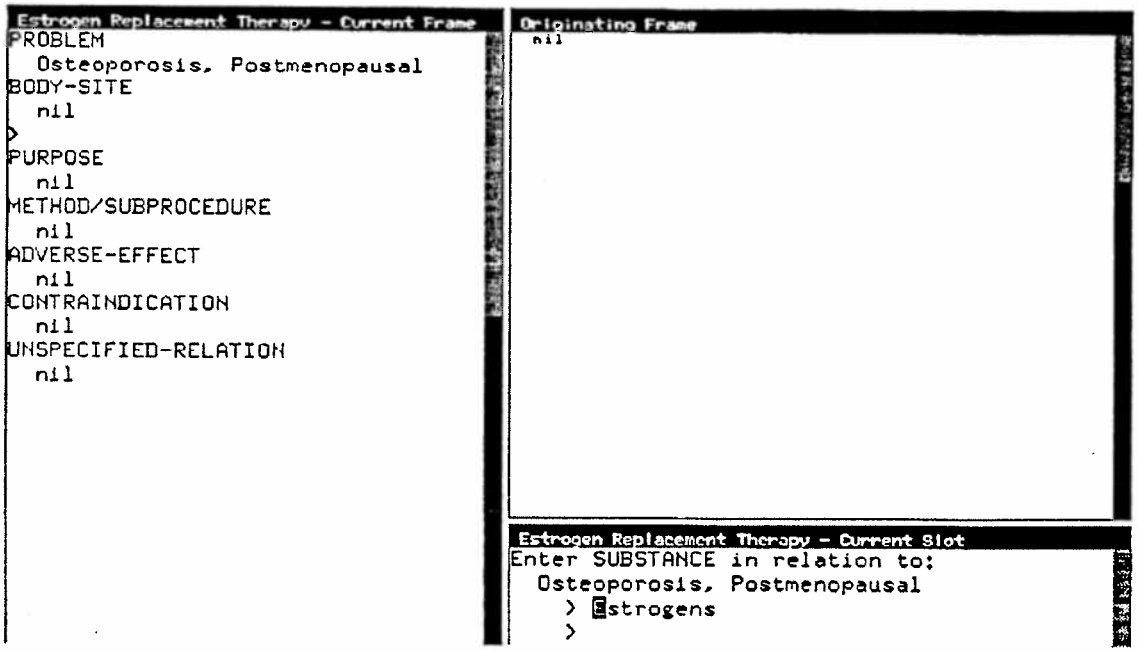

Figure 2. Indexing frame showing second slot as current slot, with display of system-generated candidate filler.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

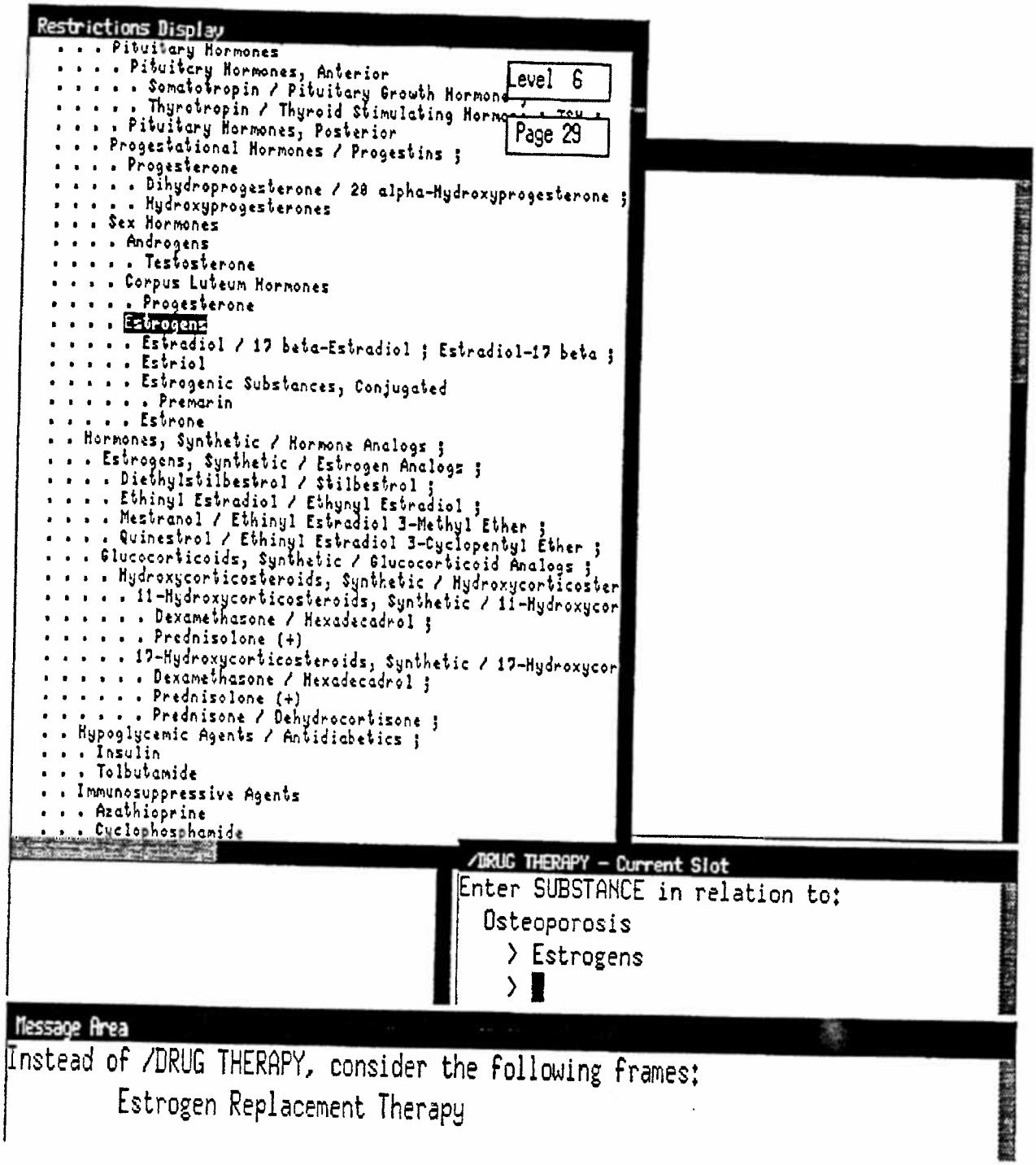

Figure 3. Indexing frame showing prescribed specificity message and restrictions display, with selection of current slot filler by mouse.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

Figure 3 shows a form of assistance that may be requested by the indexer. Here the SUBSTANCE slot is being filled in a /DRUG THERAPY indexing frame (this slash-uppercase form is expected to have a useful significance for indexers beginning to use the system). At any current-slot prompt, the indexer may request a display of possible fillers, as shown in this figure. This display is searchable (according to *regular expression* searching conventions), browsable, and mousable; as seen here, mousing on Estrogens in the display has caused this term to be entered as the current filler. This display comprises the so-called *restrictions* on the slot. Restrictions are also used for internal checking that a keyed-in filler is valid for the slot.

As a result of this Estrogens entry, the system has displayed a message bringing to the indexer's attention Estrogen Replacement Therapy, a term more specific than the current frame-term. Again, since estrogens may be administered therapeutically, but not necessarily as replacement therapy, this suggestion may be ignored. If this suggestion were taken, the indexer might cancel the current frame and begin with the suggested term. In general, when this is done, fillers of the canceled frame are automatically presented as candidate fillers for corresponding slots in the more specific frame, so that data already entered are not lost.

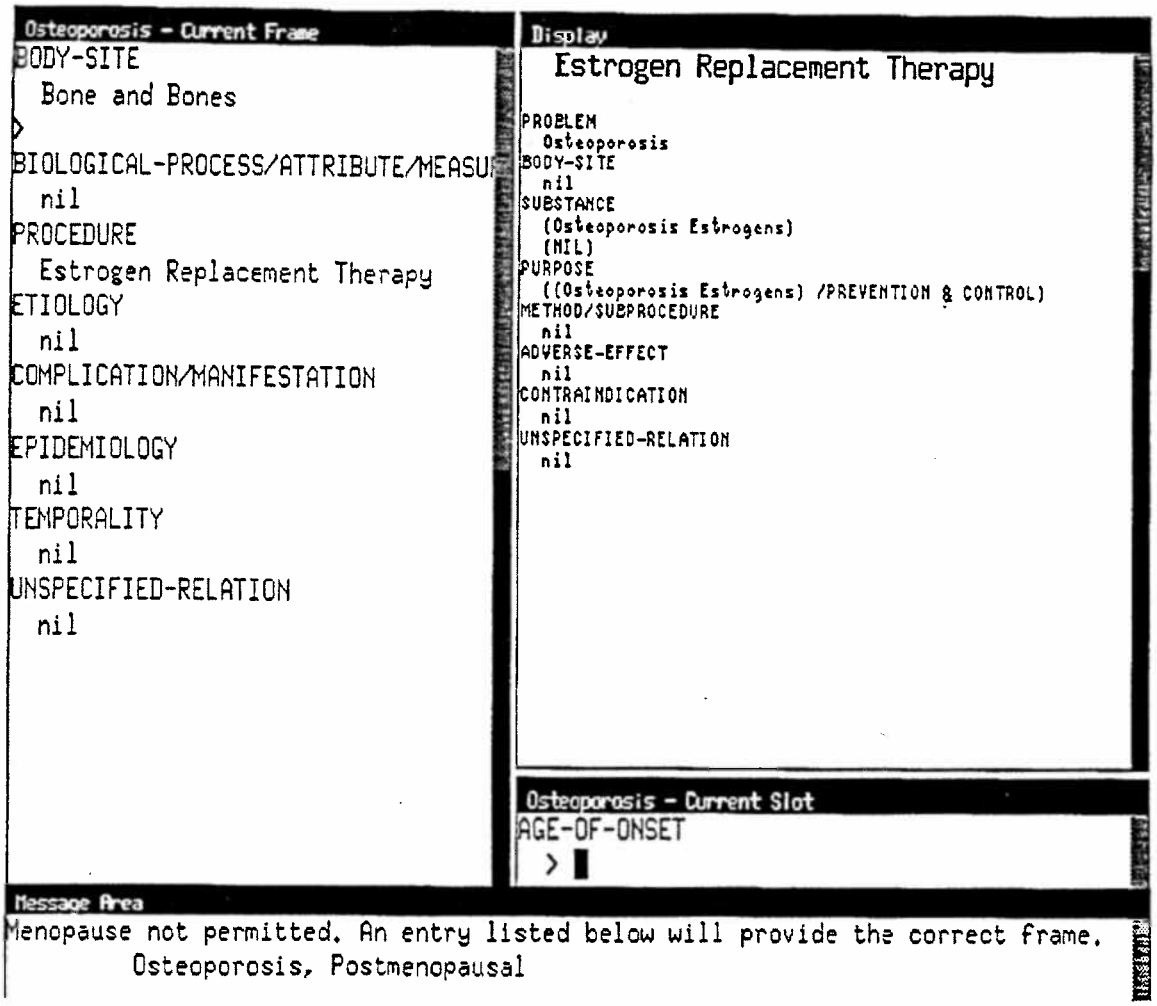

Figure 4. Indexing frame showing prescribed specificity message, and previous value that has been retrieved from stored indexing frame.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

Figure 4 illustrates a prescribed specificity. In this Osteoporosis indexing frame the indexer had entered Menopause as a filler for AGE-OF-ONSET. The system subsequently erased this filler. As indicated by the system message, this relationship requires that the more specific Osteoporosis, Postmenopausal frame be used instead of the current frame.

Another important feature is *internal retrieval.* That is, in order to minimize unnecessary data entry, the system can supply values for certain slots by fetching them from previously-filled slots in stored indexing frames. For instance, the stored frame Estrogen Replacement Therapy, as shown in the Display window in Figure 4, had been filled previously, resulting in Osteoporosis in the PROBLEM slot. Therefore, in the current frame Osteoporosis, the system has retrieved the Estrogen Replacement Therapy value for the PROCEDURE slot and then displayed it on the screen. This value is not physically in this slot,

Figure 5 illustrates system generation of conventional indexing, showing the Indices Display at some point in the indexing of a document. For instance, an indexer's adding to the ADMINISTRATION/DOSAGE slot in an Estrogens indexing frame the four fillers Administration, Oral; Drug Administration Schedule; Drug Implants; and Injections, Intravenous resulted in these four terms being added to this list of conventional indexing terms. Adding the first of these fillers caused the system to add Estrogens to this list and append to it the subheading ADMINISTRATION & DOSAGE. A Pharmacology frame with the value Estrogens in the EFFECTOR slot, and Bone Density in the EFFECT-ON slot, resulted in appending the subheading PHARMACOLOGY to Estrogens and adding Bone Density/DRUG EFFECTS. The term Pharmacology had been the value for the BIOLOGICAL-ACTION slot in the Estrogens indexing frame. Menopause was the AGE-OF-ONSET value in the Coronary Disease frame, and in this same frame, Mortality was the value for the EPIDEMIOLOGY slot. As can be seen by this example, the system is programmed to autonomously append subheadings to indexing terms based on data in these completed indexing frames.

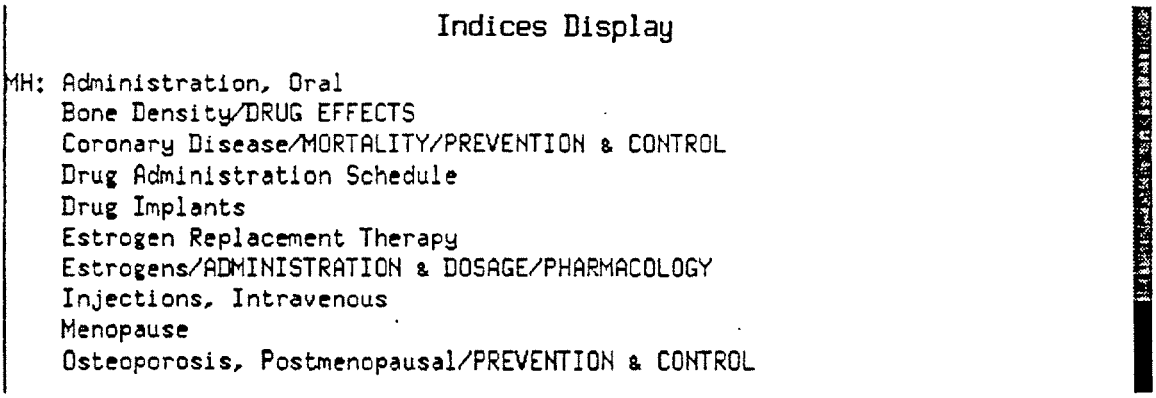

Figure 5. Indices Display showing system-generated conventional MeSH indexing corresponding to indexing frames processed so far.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

Based on the foregoing system description, it should be evident that several suggestive strategies might have been used in MedIndEx to lead to Osteoporosis, Postmenopausal. For instance, the system has rules that would suggest this term, based on indexers' linking Estrogen Replacement Therapy with either Estrogens or Menopause, in conjunction with linking one of the latter with Bone Density. Or perhaps merely the PROBLEM slot prompt in the Estrogen Replacement Therapy frame would have triggered entering the missing term.

This example also serves to illustrate how MedIndEx might provide interactive assistance to searchers. Given the search query emphasis on osteoporosis, not mentioning the concept of bone density, if the system were to index the query, it might well suggest the search term Bone Density, thereby assisting the searcher in compensating for the missing osteoporosis indexing term. Furthermore, even with ideal indexing, infrequent searchers may not be aware of pre-coordinate indexing terms like Estrogen Replacement Therapy and Osteoporosis, Postmenopausal. In filling a query frame with simple terms like Estrogens and Osteoporosis, based on linkages in user-created query frames and rules in the KB that use these relationships, the system would be able to suggest these terms. Although a premise of this project has been that beginning with the indexing problem, which is centralized at NLM, is a reasonable start in applying AI techniques to information retrieval, now that we have developed a knowledge base and software that might be adapted for helping with searching directly, it seems also reasonable, not to mention intriguing, to conduct at least a preliminarily investigation of this additional use of MedIndEx. We are currently experimenting with this.

4. KB MANAGER

Managing the KB, that is, creating and editing knowledge-base frames, is more complicated than managing a thesaurus, since the knowledge engineer (formerly known as the thesaurus specialist) is responsible not only for the terms, but also encoding data and procedures needed for providing interactive indexing assistance. The system, in effect, merges a thesaurus and indexing manual, in a potentially concise, executable form.

In particular, it would be impossible to manage a sizeable KB that uses inheritance without a tool like KB Manager. Frames often contain procedures that access information (i.e., data or other procedures) from other frames. While this helps ensure consistency (and reduces redundancy), this benefit may be somewhat counteracted by the difficulty in determining whether new information in a frame is consistent with existing information found in frames quite remote from the current frame. This remoteness may be vertical (inheriting from several nodes above the current frame) or horizontal (accessing frames along non-inheritance links). This situation is made even more difficult in cases of multiple inheritance.

General requirements of a knowledge base are that it be consistent and have proper syntax. Modifying the KB is essentially the process of editing facet-fillers (contents of facets). General functions performed using KB Manager, a tool we have developed, are summarized as follows:

- Creating new frames, by adding frame terms to CHILDREN slot, VALUE facet, of existing frames.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

- Making inheritance links, by adding frame terms to INHERITS-FROM slot, VALUE facet, of lower-level frames.

- Encoding indexing assistance, by modifying fillers of facets, such as RESTRICTIONS, IF-NEEDED, CAN-CONTINUE?, and so forth.

KB Manager automatically checks for inconsistencies and requires their resolution before the user may proceed. The system accesses inherited information, evaluates it, uses it for checking new information, and also displays inherited and local information for selection of preferred inheritance paths or for overriding inheritance with local information. A menu interface for selecting slot and facet names ensures proper syntax down through the facet name.

Special interfaces have been developed, not only for consistency and proper syntax, but also for ease of programming. It may be possible for software to be managed by people who are not necessarily expert programmers. For instance, the system evaluates code to display hierarchies in the form of menus, and conversely analyzes the selections a user makes by mousing, and automatically generates code corresponding to these selections. Selections are treated as a flat list, which is then matched to the code for the complete, master hierarchy. Thus, selections of the same terms from menus corresponding to any segment of the hierarchy will always result in consistent code and therefore consistent displays. Another interface style we use is direct manipulation of code as objects. This method of building facet contents uses menus consisting of code, cut & paste, and syntax-checking of code that has been built. This type of interface is made possible by public-domain software known as CLOS (Common Lisp Object System). Finally, we also offer as a standard option the use of the system text editor which checks syntax of computer code.

5. HARDWARE AND SOFTWARE

The prototype is written in Sun™ Common Lisp 3.0 and runs on the SPARCstation 1™ workstation under the SunOS® operating system. This machine has 28 Mbytes of memory and a 327-Mbyte disk. Currently, two interfaces are being maintained. The original workstation interface uses SunView™-based Window Tool Kit (a library of Lisp functions). Recently, for portability, we have developed an X Window interface using X11 Release 4. In addition to CLOS, mentioned earlier, we also use CLX and CLUE. The system can be run in an architecture with two ethernetted SPARCstations, one as client, the other as X server. We also have experimented with access to the prototype from a PC ethernetted to the workstation using X server software on the PC.

Domain-specific code is about 1.8 Mbytes, including 1.3 Mbytes for the actual frames, with most of the rest for word and term aliases and term sort versions. There are more than 3200 frames (MeSH concepts), with 43 slots that may be filled in the indexing system, and from 1-12 slots for any one frame. The project has generated about 1.1 additional Mbytes of code for running the MedIndEx application, but this code is essentially domain-independent. We are designing the system so that it might be used for similar applications, but in other domains. This aspect of our design has been successfully demonstrated using the AAT (Art and Architecture Thesaurus) and a

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

few documents from the Art Literature International (RILA) database in a quick experiment (test documents were on French Gothic cathedrals).

6. TRIAL USE AND EVALUATION

We are currently preparing the system for trial use by indexers, and designing an evaluation. There are problems in rigorously evaluating incomplete knowledge-based systems, although it may be possible to conduct a formal evaluation using a medical subdomain. On the other hand, a more qualitative evaluation providing feedback from indexers using MedIndEx seems of paramount importance at this time. Questions in the following two areas would be addressed:

- Organization, representation, and presentation of procedural knowledge to assist users performing the intellectual task of indexing.

- The effect of knowledge-based systems on indexing output.

Specific questions include: Is the level of granularity of the knowledge base appropriate for the indexing application? Are the relations subdividing the concepts in the knowledge base complete and non-overlapping (relations are used to prompt indexers for information)? Do the principles of inheritance the system uses (whereby a concept automatically accesses the relations, procedures and facts associated with a parent concept in the inheritance-based classification) work gracefully, in particular multiple inheritance? How do users react to transferring to a system that uses knowledge-based guidance from one where this guidance is absent. What are the rate-limiting steps during system use? Is it easy for users to interact with the system? What facilities and capabilities do users need? Is the system-generated conventional indexing output consistent with expert indexing? Does the system result in inter-indexer consistency? Does knowledge-based indexing result in more exhaustive indexing than conventional indexing?

In particular, we recognize measuring indexing quality as problematic. O'Connor wrote that indexing quality is appropriately determined by measuring retrieval quality and that an indexing duplication study (comparing mechanized indexing to human indexing corrected for errors) should not be called a test of mechanized indexing methods, but rather thought of as a way to investigate such methods empirically and probably less expensively than retrieval testing [7]. Although his point is well-taken, we think there is no other choice but to measure indexing "quality" in terms of traditional indicators of inter-indexer consistency and expert indexing duplication using scoring methods.

In addition, given that even ideal indexing is focused on the document, rather than all possible search queries the document might answer, we will further define our investigation of MedIndEx as a search assistant that provides knowledge-based semantic associations useful particularly for retrieval, pending our preliminary work this year. For this, we might use the set of queries selected for the NLM test collection, in particular some of the fifty-one queries in clinical medicine which retrieved 1,262 citations which then were assessed for relevance [8].

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

7. REFERENCES

[1] ***National Library of Medicine programs and services fiscal year 1989.*** Bethesda, MD, 1990.

[2] **Humphrey, S.M. and Chien, D.** *The MedIndEx System: research on interactive knowledge-based indexing and knowledge management.* Technical Report NLM-LHC-90-03. Bethesda, MD: National Library of Medicine, July 1990. Distributed by NTIS: Publication No. PB90-234964/AS.

[3] **Lancaster, F.W.** *Indexing and abstracting in theory and practice.*Champaign, IL: University of Illinois Graduate School of Library and Information Science, 1991.

[4] **Hutchins, W.J.** "Linguistic processes in the indexing and retrieval of documents." *Linguistics* 1970 Sep;61:29-64.

[5] **Slamecka, V.** "Classificatory, alphabetical, and associative schedules as aids in coordinate indexing." *American Documentation* 1963 Jul;14(3):223-8.

[6] **Humphrey, S.M.** "MedIndEx: The Medical Indexing Expert System." Chapter in: *Expert Systems in Libraries*, eds. Rao Aluri and Donald E. Riggs. Norwood, NJ: Ablex, 1990.

[7] **O'Connor, J.** Mechanized indexing methods and their testing. *Journal of the Association for Computing Machinery* 1964;11(4):437-49.

[8] **Schuyler, Peri L., McCray, Alexa T., and Schoolman, Harold M.** "A test collection for experimentation in bibliographic retrieval." Paper in: *MEDINFO 89: Proceedings of the Sixth Conference on Medical Informatics*. Amsterdam: North-Holland, 1989; 910-2.

ISSN: 2324-9773

Humphrey, S. (1991). Use and Management of Classification Systems for Knowledge-Based Indexing. 2nd ASIS SIG/CR Classification Research Workshop, 49-62. doi:10.7152/acro.v2i1.12547

ISSN: 2324-9773