Analysis of a Study of the Users, Uses and Future Agenda of the UMLS

Yan Chen, MS[1,3], Yehoshua Perl, PhD[1], James Geller, PhD[1], James J. Cimino, MD[2]

[1]Department of Computer Science, New Jersey Institute of Technology

[2]Department of Medical Informatics, Columbia University

[3] CIS Department, Borough of Manhattan Community College, CUNY

Reprint Requests & All Communications:

Yehoshua Perl

perl@oak.njit.edu

Department of Computer Science

New Jersey Institute of Technology

University Heights

Newark, NJ 07102

Tel: (973) 596-3392

Fax: (973) 642-7029

ABSTRACT

Objective: The UMLS constitutes the largest existing collection of medical terms. However, little has been published about the users and uses of the UMLS. This study sheds lights on these issues.

Design: We designed a questionnaire consisting of 26 questions and distributed it to the UMLS user mailing list. Participants were assured complete confidentiality of their replies. To further encourage list members to respond, we promised to provide them with early results prior to publication. Sector analysis of the responses, according to employment organizations is used to obtain insights into some responses.

Result: We received 70 responses. The study confirms two intended uses of the UMLS, access to source terminologies (75%) and mapping among them (44%). However, most access is just to a few sources, led by SNOMED, MeSH and ICD. Out of 119 reported purposes of use, terminology research (37), information retrieval (19), and terminology translation (14) lead. Four important observations are that the UMLS is widely used as a terminology (77%), even though it was not designed as one; many users (73%) want the NLM to mark concepts with multiple parents in an indented hierarchy and to derive a terminology from the UMLS (73%). Finally, auditing the UMLS is a top budget priority (35%) for users.

Conclusion: The study reports many uses of the UMLS in a variety of subjects from terminology research to decision support and phenotyping. The study confirms that the UMLS is used to access its source terminologies and to map among them. Two primary concerns of the existing user base are auditing the UMLS and the design of a UMLS-based derived terminology.

Keywords: Unified Medical Language System (UMLS), UMLS Users, UMLS Uses, Future of the UMLS, UMLS Auditing, Terminology, Source Terminologies.

I INTRODUCTION

The National Library of Medicine (NLM) is the sponsor of the Unified Medicine Language System (UMLS) [1, 2, 3] project, which is an effort to overcome the fundamental barriers of communication and the lack of a standard machine-readable language in medicine [3]. The UMLS has a large user population. Users are licensed by NLM and are supported by a UMLS user mailing list [4]. The UMLS is supposed to represent the conceptual connections between user questions and the effective retrieval of relevant machine-readable biomedical information. The UMLS is a set of machine-readable knowledge sources, including the Metathesaurus [5, 6] which contains 1.3 million concepts derived from a variety of more than 100 existing biomedical vocabularies and classifications, the Semantic Network [7] representing Semantic Types and sensible relationships among them and the SPECIALIST lexicon, which provides the lexical information needed for the SPECIALIST Natural Language Processing System [8]. The UMLS was proposed with the expectation of supporting interface programs that use these UMLS knowledge sources to emulate some of the functions of an expert reference librarian, while accessing a broad range of information. According to Humphreys et al. [2], the basic uses of the UMLS are the controlled vocabulary function of the Metathesaurus and the Semantic Network, the enhanced information retrieval from various sources and mapping among them and support for Natural Language Processing.

Over the past two decades, the UMLS has been developed steadily, increasing in size to become the largest existing collection of medical terms. From a computer science perspective in the context of ontologies, the UMLS is viewed in [9, 10] as a well developed, usable, very large ontology with a long lifespan, which is used in multiple projects. It has been distributed in many healthcare organizations. Every year a substantial number of papers are being published about the UMLS, e.g. [11, 12, 13]. However, it is difficult to find in the literature what the UMLS is actually used for, by whom and how. Has it been used as initially expected? Which features proved useful? Which extra features are desired by users?

Hollis [14] addressed some of these issues briefly in her UMLS study. Hollis' seven questions required free-text answers. None of her questions were about expectations and possible improvements of the UMLS. As in our study, Hollis' survey was distributed through the UMLS user mailing list [4]. Only ten responses were received by Hollis, which makes it hard to discern UMLS users' general opinions. In this paper we present a study that sheds more light on the user population of the UMLS, the way they use it and their preferences.

In particular we are interested in examining the intended use of the UMLS versus actual use reported by the respondents. For example, the UMLS was designed as an ontology supporting access to and mappings among over 100 medical source terminologies [2]. To what extent is the UMLS used for these purposes?

Furthermore, effective with the first 2004 release, the Metathesaurus's Rich Release Format (RRF)[15] represents sources "transparently." That is, both users and applications can access its source vocabularies' content without loss of information. The concept-based abstractions in the Original Release Format (ORF) prevent the perfect extraction of a few sources because of differences between the Metathesaurus concept-based representation and the code-based nature of these sources. This small loss of information has been eliminated in RRF so that, for instance, the distinction between a source's inter-term or inter-code relationships and the information added in the creation of the Metathesaurus is made [16]. Do users utilize the transparent access to sources? Which sources are accessed most? Which subject areas of the UMLS are extensively used and for which areas do UMLS users wish to extend the coverage? The UMLS was not designed to be a terminology. However, following messages posted to the UMLS user mailing list [4], we have encountered anecdotal evidence of UMLS use as if it were a terminology. What percentages of the users use the UMLS as a terminology? If this percentage is indeed high, as we had hypothesized [17], would those users like the NLM to design a terminology derived from the UMLS? In this derived terminology the information

about the occurrence of terms in source terminologies would be removed. Thus, inconsistencies found in the UMLS could be removed from the derived terminology. Note that information regarding occurrence in source terminologies will still be available to a user in the UMLS, which will not change. For more ideas for such a terminology design, see [17].

Due to the UMLS being integrated from many source terminologies as well as its size and complexity, it is unavoidable that some classification errors and inconsistencies have been introduced. In recent years we have seen a surge in publications discussing techniques for auditing the UMLS, e.g. [18-26]. Do users care about errors in the UMLS and which ones concern them most?

We examined users' opinions regarding two interface features, one offered by NLM and the other suggested in the survey. To get a deeper understanding of the results, we used sector analysis for some questions. At last, we asked users what percentages of a putative UMLS budget should be allocated to different tasks. In this report, we present these issues with the intention of providing the NLM and UMLS users with constructive feedback regarding future potential development of the UMLS. This study was neither initiated nor supported by the NLM to assure its independence.

II METHODS

We designed a 26-question survey, consisting of three parts. The first part was about demographics and employment information of UMLS users. The second part was about various aspects of use of the UMLS. The last set of questions concerned the "UMLS agenda" and challenged the users to express their priorities. The third part, which is more complex and includes explanations, appears in Appendix II. For the whole questionnaire, see www.cis.njit.edu/~oohvr/new/umlsstudy.doc. For some questions, for example about professions, highest educational degrees, mode of use and kind of host systems, multiple responses were allowed.

We sent our questionnaire to the UMLS user mailing list maintained by the NLM [4]. This mailing list had about 600 members at that time. Participants were assured complete confidentiality of their replies. To further encourage list members to respond, we promised to provide them with early results prior to publication. After an initial tight deadline, we extended the deadline and sent reminders.

As opposed to Hollis' UMLS study, we kept the number of open-ended questions to a minimum in order to minimize the efforts and response time of respondents and so increase their number. All but four questions allowed the respondents only to choose among a few given options, although a choice "Other" was given. For the four questions, we did not want to bias the respondents by our choices, and allowed them to enter their own choices.[1]

For some questions, we use sector analysis, where we distribute the answers over the employment types of the respondents, to gain better insights into the responses.

We use tables to display the absolute values (and in parenthesis the percentages) of the options for the various questions. This duality of the numerical data is especially helpful in cases of multiple answers where the percentages add up to more than 100%. For displaying results of a sector analysis, both segmented (stacked) bar charts [27] and tables are used. They help to visually highlight the options where some sectors display a digression from the results for the overall population of the study. Furthermore, to emphasize popular options, the different subject areas in a segmented bar chart are listed in decreasing order of percentages, upwards, according to the distribution of the overall study population. Examples of such digressions are analyzed in the Discussion Section. The tables are presented for the readers who prefer numerical information over graphical representation.

III RESULTS

There were 70 respondents for our questionnaire. A 50% increase of initial submissions was achieved by sending a reminder and extending the deadline.

A Demographics and Employment

The majority of respondents, 70%, are from the USA, followed by 20% from Europe, 4.3% from Canada and the rest from other continents**.** Table A1 (in the Appendix I) shows users' age distribution. The largest age group is 51-60. The users' highest education level is shown in Table 1. About 21% of users have 2 highest degrees (i.e., not in the same field), out of which, 9% have PhD and MD degrees, 9% have Master's and MD and 3% have Master's and PhD degrees. Hence, among those who have 2 degrees, 86% have MD degrees. The average number of degrees reported per user is 1.17.

Table 1: Highest Education Level

Degree	Number (%)
Master	27 (39%)
Ph.D.	23 (33%)
M.D.	23 (33%)
Bachelor	5 (7%)
Others	3 (4%)

Table 2: Professions

Profession	Number (%)
Researcher	33 (47%)
Programmer	16 (23%)
Physician	15 (21%)
Professor	12 (17%)
Student	11 (16%)
Manager	8 (11%)
Nurse	5 (7%)
Librarian	4 (6%)
Administrator	4 (6%)
Engineer	3 (4%)
Other	12 (17%)

Table 3: Employment Sectors

Organization	Number (%)
University	34 (49%)
Industry	25 (36%)
Government	6 (9%)
Self-Employed	2 (3%)
Research Institution	1 (1%)
Others	2 (3%)

Table 4: Industry (36%, see Table 3) Sectors

Employer Type	Number (%)
SW Vendor	11 (44%)
Hospital	6 (24%)
Health Information Processing	3 (12%)
Insurance	2 (8%)
Doctor's Office	1 (4%)
Pharmaceutical Company	0 (0%)
Others	2 (8%)

The first employment question was on users' professions, as shown in Table 2. On average, a user listed 1.77 professions, with 37% listing multiple professions. About 23% listed 2 professions. One respondent listed 5 professions: engineer,

manager, professor, programmer and researcher. We further asked for the employment sectors in which they are active (see Table 3). Table 4 shows the sector distribution of the 36% employed in industry. About half and a quarter of industry employees are from software vendors and hospitals, respectively. As for the organization size, more than half of the organizations, mainly universities, have over 1000 employees (see Table A2).

B The Uses of the UMLS

1) Length of Experience with UMLS

More than 1/3 (37%) of users have used the UMLS less than a year, 24% and 22% have used the UMLS 2~4 years and 5~7 years, respectively, while 17% have been users for 8 years or more.

2) Subject Areas

Figure 1: Percentages of Subject Areas Usages by Organizations

We listed 15 subject areas in the questionnaire, based on the partition of the Semantic Network in the NLM UMLS website [28]. In Figure 1 we use segmented bar charts to show the subject area percentages of use for the overall study population and for the population of three kinds of employment organizations. On average, respondents reported using 5.7 areas and a subject area is used by 47% of the users. The sector analysis shown in Figure 1 and Table 5 is intended to distinguish the use of the subject areas among universities, software companies

and healthcare organizations compared to the overall results. (Software companies consist of Software Vendors, Health Information Processing companies and Insurance companies while healthcare organizations consist of Hospitals and Doctor's Offices.) The combination of similar employment organizations into the broader groups of software companies and healthcare organizations helps to obtain a clearer picture of the distribution by combining small groups like doctor's offices into larger similar groups, like hospitals. Table A3 shows the distribution of the user numbers according to the number of areas selected. Only seven users were interested in all 15 areas. The leading subject areas with over 60% interest across all sectors are Concepts and Ideas, Disorders and Procedures.

Ta**ble 5: Percentages of Subject Areas Usages by Organizations**

	Overall	**University**	**Software**	**Healthcare**
Activities & Behaviors	34	32	53	30
Anatomy	54	62	53	60
Chemicals & Drugs	51	38	71	70
Concepts & Ideas	67	65	59	70
Devices	36	29	53	50
Disorders	63	59	65	90
Geographic Areas	14	15	24	10
Genes & Molecular Sequences	33	29	41	50
Living Beings	30	32	35	20
Objects	26	26	35	10
Occupations	19	18	29	30
Organizations	16	15	18	20
Phenomena	20	21	18	10
Physiology	43	25	53	70
Procedures	60	59	71	60
Average	38	36	45	43

3) Mode of Operation

Figure 2 uses segmented bar charts to show the sector analysis for the modes of operations. Note that the percentages in each segmented bar add up to more than 100% since some users use the UMLS in multiple modes. The same

phenomena appear for all three issues for which we conducted sector analysis (Figures 1, 2.and 3). Table 6 shows the same information numerically.

Figure 2: Percentages of Modes of Operation by Organizations

Table 6: Percentages of Modes of Operation by Organizations

	Overall	University	Software	Healthcare
Research	73	74	76	60
Prototype Design	31	38	29	30
Testing	23	21	24	30
Production	17	15	18	40
Other	11	9	6	10
Average	31	31	31	34

Table 7 : Host Systems for the U

Host Systems	Number (
Medical Research	35 (5
Terminological	28 (4
Clinical Information	28 (4
Decision Support	23 (3
Billing	3 (
Others	12 (1

Table 8: Original Response Examples

Category (Total Number in Category)	Selected Example Original Text in this Category
Terminology Research (37)	General terminology browsing
	Assure consistent use of terminology
	Procedure names
	Building medical ontologies
	Source of synonyms
	Foundation for Vocabulary Management and natural language processing of medical standard terminology for public health integrated systems
	Semantic network extract and modeling
	Provide concept search interface
Information Retrieval (19)	Data content searches
	Conceptual text indexing
	Text mining of biological literature
	Indexing medical documents
Terminology Translation (14)	Mapping concepts across vocabularies
	Match own terminology to commonly accepted codes
	Multilingual to English translations for queries to Entrez-PubMed
UMLS Research (13)	Auditing the UMLS
	Study mappings between terminological systems
	Study of Metathesaurus structure using Complex Network theory.
	Research on terminology coverage
Electronic Health Record (10)	Building problem lists using UMLS concepts
	Providing a unified method for storing information and knowledge in an EMR
	Relate UMLS vocabularies to EHR models associated attributes
Natural Language Processing (8)	Collecting linguistic knowledge for Natural Language Processing
	Parsing medical abstracts
	Create lexicon files for NLP
Education (5)	Building an ontology for a medical education system
	Educational resource for Informatics program
Decision Support (3)	Decision support modeling
	Mark up oncology guidelines
System Development (3)	Creating a clinical trails scheduling system
	Development of a speech ordering system for tests, meds, imaging, etc.
Billing (1)	Drug-disease linkages for billing purposes

Category (Total Number in Category)	Selected Example Original Text in this Category
Terminology Research (37)	General terminology browsing
	Assure consistent use of terminology
	Procedure names
	Building medical ontologies
	Source of synonyms
	Foundation for Vocabulary Management and natural language processing of medical standard terminology for public health integrated systems
	Semantic network extract and modeling
	Provide concept search interface
Information Retrieval (19)	Data content searches
	Conceptual text indexing
	Text mining of biological literature
	Indexing medical documents
Terminology Translation (14)	Mapping concepts across vocabularies
	Match own terminology to commonly accepted codes
	Multilingual to English translations for queries to Entrez-PubMed
UMLS Research (13)	Auditing the UMLS
	Study mappings between terminological systems
	Study of Metathesaurus structure using Complex Network theory.
	Research on terminology coverage
Electronic Health Record (10)	Building problem lists using UMLS concepts
	Providing a unified method for storing information and knowledge in an EMR
	Relate UMLS vocabularies to EHR models associated attributes
Natural Language Processing (8)	Collecting linguistic knowledge for Natural Language Processing
	Parsing medical abstracts
	Create lexicon files for NLP
Education (5)	Building an ontology for a medical education system
	Educational resource for Informatics program
Decision Support (3)	Decision support modeling
	Mark up oncology guidelines
System Development (3)	Creating a clinical trails scheduling system
	Development of a speech ordering system for tests, meds, imaging, etc.
Definitions (1)	Source of Definitions and cross-maps used on Diseases Database Website
Adverse Events (1)	Coding Adverse event terms for clinical trials

Category (Total Number in Category)	Selected Example Original Text in this Category
Terminology Research (37)	General terminology browsing
	Assure consistent use of terminology
	Procedure names
	Building medical ontologies
	Source of synonyms
	Foundation for Vocabulary Management and natural language processing of medical standard terminology for public health integrated systems
	Semantic network extract and modeling
	Provide concept search interface
Information Retrieval (19)	Data content searches
	Conceptual text indexing
	Text mining of biological literature
	Indexing medical documents
Terminology Translation (14)	Mapping concepts across vocabularies
	Match own terminology to commonly accepted codes
	Multilingual to English translations for queries to Entrez-PubMed
UMLS Research (13)	Auditing the UMLS
	Study mappings between terminological systems
	Study of Metathesaurus structure using Complex Network theory.
	Research on terminology coverage
Electronic Health Record (10)	Building problem lists using UMLS concepts
	Providing a unified method for storing information and knowledge in an EMR
	Relate UMLS vocabularies to EHR models associated attributes
Natural Language Processing (8)	Collecting linguistic knowledge for Natural Language Processing
	Parsing medical abstracts
	Create lexicon files for NLP
Education (5)	Building an ontology for a medical education system
	Educational resource for Informatics program
Decision Support (3)	Decision support modeling
	Mark up oncology guidelines
System Development (3)	Creating a clinical trails scheduling system
	Development of a speech ordering system for tests, meds, imaging, etc.
Knowledge Management (1)	Storing, presentation and processing of knowledge
Phenotyping (1)	Phenotyping

4) Host Systems for UMLS Use

The UMLS is a collection of vocabularies and tools. It is not, by itself, a working software system. However, programmers may incorporate the information contained in the UMLS into other fully functional software systems. We refer to such a software system as the "host system" of the UMLS. Table 7 shows the kinds of host systems in which the UMLS is used. An average of 1.9 kinds per respondent was reported. Half of the users use the UMLS in Medical Research Systems. The trailing host systems are Clinical Information Systems and Terminological Systems both with 40%.

5) Purposes of UMLS Use

Figure 3: Percentages of Purpose of UMLS Uses by Organizations

Many users reported multiple uses of the UMLS, resulting in a total of 119 uses.[1] Table 8 gives several examples of the original responses for each category. In Figure 3, (Table 9) segmented bar charts show sector analysis of categories of

use by organizations, listed in percentages. Note that the percentages in each segmented bar are adding up to more than 100% since users listed multiple uses. Table 9 shows the sector analysis in a numerical format.

Table 9: Percentages of Purpose of UMLS Uses by Organizations

	Overall	University	Software	Healthcare
Education	7	12	0	10
Electronic Health Record	14	18	6	30
Information Retrieval	27	26	35	10
Natural Language Processing	17	24	0	0
UMLS Research	19	32	0	10
System Development	4	6	0	0
Terminology Research	53	35	71	80
Terminology Translation	20	18	35	0
Decision Support	4	0	6	20
Special Uses	4	6	6	0
No answers	16	15	24	10
Average	17	17	17	15

6) Access to UMLS Source Terminologies

We found that 80% of users access the UMLS source terminologies, such as CPT (Current Procedural Terminology), ICD (International Statistical Classification of Diseases and Related Health Problems), MeSH (Medical Subject Headings) , SNOMED (Systematic Nomenclature of Medicine), NDF-RT (National Drug File - Reference Terminology), etc. About four fifths of answers from 104 responses[1] are made up of just four terminologies, SNOMED (32), MeSH (23), ICD (21) and CPT (8), while LOINC (Logical Observation Identifier Names and Codes) (4) and RxNORM (a standard terminology for drug products) (3) follow. Other terminologies mentioned twice are NANDA (North American Nursing Diagnosis Association Taxonomy), NCI (National Cancer Institute Thesaurus), NIC (Nursing Interventions

Classification), and mentioned once are ALT (Alternative Billing Concepts), DSM-IV (Diagnostic and Statistical Manual of Mental Disorders), GO (Gene Ontology) , HL7 (Health Level Seven Vocabulary), MED (Medical Entities Dictionary), RUS2002 (Russian Translation of MeSH) and UWDA (University of Washington Digital Anatomist).

Effective with the first 2004 release, the Metathesaurus' Rich Release Format represents sources "transparently." Only 20% are using the UMLS transparency feature, while 56% of the respondents plan to use this feature in the future.

7) Using the UMLS as a Mapping Tool

About 44% of the respondents are using the UMLS as an ontology, supporting mappings between its various source terminologies [2].

8) Use of the UMLS as a Terminology

A full 77% of respondents verified that they are using the UMLS as a terminology, even though it was not designed to be a terminology.

C AGENDA

1) Derived Terminology

Almost 73% of respondents stated that they would like the NLM to design a terminology derived from the UMLS.

2) Expanding the UMLS Coverage

We found that users would like the NLM to expand the UMLS coverage in 25 areas[1]. The leading requested areas are Genomics, Biology and Finding with 6, 5 and 4 respondents respectively. Three users requested the following areas: Drugs, Mapping, Globalization, Procedures, Signs and Symptoms,

Sociology and Therapy, while two requested Coding Systems, Diseases and Disorders. Eleven other areas, not listed, were from just one respondent.

3) Modeling Errors

Table 10 shows the average level of (users) concerned with different kinds of modeling errors. Six kinds of modeling errors were offered, namely concept redundancy [18], concept polysemy (also called ambiguity) [19], wrong hierarchical relationships [19, 20], wrong associative relationships, wrong semantic type assignments [21, 22] and redundant semantic type assignments [23, 24]. We listed "not at all," "a little," "moderately" and "a lot" as the choices indicating the level of concern. When analyzing the data, we assigned an integer score 0, 1, 2 or 3 to each choice, respectively. The average concern level for all modeling errors is 1.72. The leading errors for which users are moderately concerned are wrong semantic type assignment and wrong associative relationships.

4) Missing Terminological Knowledge Elements

Table 11 shows the average concern level for missing knowledge elements such as missing concepts [25, 26], missing definitions, missing synonyms [26], missing hierarchical relationships [25, 26], missing associative relationships [26] and missing semantic type assignments [25]. The average concern level for all missing knowledge elements is 1.57. The combined average concern level for both wrong and missing knowledge elements is 1.65.

Table 10: Average Concern Levels about Modeling Errors
(3 =A lot, 2 = Moderately, 1= A little, 0 = Not at all)

Wrong Semantic Type Assignments	2.14
Wrong Associative Relationships	2.11
Wrong Hierarchical Relationships	1.97
Concept Redundancy	1.53
Redundant Semantic Type Assignments	1.3
Concept Polysemy	1.26

Table 11: Average Concern Levels of abou Missing Knowledge Elements
(3 =A lot, 2 = Moderately, 1= A little, 0 = No at all)

Missing Hierarchical Relationships	1.86
Missing Semantic Type Assignments	1.76
Missing Synonyms	1.51
Missing Concepts	1.45
Missing Associative Relationships	1.45
Missing Definitions	1.43

5) Interface

Our results show that 73% of participating users would want the NLM to enhance the UMLSKS META interface to mark a concept with multiple parents with a "*" in the indented hierarchy (similar to the way MS Windows marks a directory with children by a "+"), 19% chose "No." There were 74% of the respondents who wanted to see the Semantic Navigator maintained for future releases,[2] while 16% answered "No." We challenged the users to suggest other enhancements they might want to see.[1] Only 10% of the respondents answered this question and their replies varied widely. Some suggested a better interface without offering specifications. Some requests were for better integration of foreign languages, transparent mapping, and software to assist those who want to contribute groups of new/missing terms to fill gaps in the UMLS. Some answers were irrelevant.

6) Goals for UMLS Development

UMLS users expressed what percentages of a putative NLM budget for the UMLS should be allocated to Auditing, Derived Terminology Development, Expansion of New Subject Areas, Better Interfaces, and Others (Figure 4).

Figure 4: Desired Budget Allocation

IV DISCUSSION

A Demographics and Employment

The vast majority of UMLS users are from the USA (45 respondents), followed by Germany (6), Canada (3) and France (2). In spite of the efforts to supply multi-language support for the UMLS, only limited interest (30%) was shown outside of the USA. About 60% of UMLS users are above 40 years old. As a group, UMLS users are highly educated with 59% holding PhD or MD degrees. This is consistent with the result that the top profession of respondents is researcher, followed by programmer, physician, professor and student. We understand the heavy use of "researcher" to mean that some users interpret "researcher" in a broad way. This is especially true when some users considered themselves to be in multiple professions.

B Uses of the UMLS

Five of the top six areas used are biomedical subjects led by Disorders and Procedures, as expected, since the UMLS covers the medical field. The leading interest in the abstract subject of Concepts and Ideas is surprising, due to two aspects. First, the UMLS is used mainly for medical knowledge, so a top interest which is a non-medical subject is surprising in this context. Second, most users are typically interested in concrete knowledge rather than ideas and conceptual knowledge. It would be interesting to investigate which concepts of this subject users are actually interested in. The only two biomedical subjects with lower interest are (1) Devices (2) Genes & Molecular Sequences, but the interest in them is higher for software and healthcare employees. Healthcare employees lead the interest in Disorders, Physiology, Concepts and Ideas, and Genes & Molecular Sequences. University employees lead in Anatomy and Phenomena. Software companies lead in interest in all other subjects and are especially interested in Procedures as well as Chemicals and Drugs.

We found that in all sectors, the UMLS mode of operation is mainly for research, by a wide margin. Of special interest is that overall 17% report using the UMLS in production, where healthcare employee report 40%.

In terms of the purposes of using the UMLS, Terminology Research is the primary purpose for all of the organization categories. Universities lead in most categories, such as Education, Natural Language Processing (NLP), UMLS research and System Development; Software companies lead in Information Retrieval and Terminology Translation while Healthcare organizations use Electronic Health Record (EHR) and Terminology Research most.

Three quarters of the users access the UMLS source terminologies as originally intended. In view of this result one would expect that users are extensively utilizing the transparent access to source terminologies added in 2004. This feature enables a user to obtain, through the UMLS, a source terminology, as originally created. However, use of transparent access is spreading slowly. Only about 20% from the 75% of the users accessing source terminologies use transparent access, although more than 50% intend to use it. Perhaps more aggressive advertising by NLM, of this recent feature, will make it more popular. Software company employees lead in accessing the UMLS source terminologies transparently.

Actual access that was reported was limited to just a few source terminologies. Only 17 source terminologies were mentioned and just four, SNOMED, MeSH, ICD and CPT, accounted for 80% of the responses. Users do not seem to care much about the other source terminologies' internal representations of UMLS concepts, as we speculated in [17]. Indirectly, this evidence supports the idea of deriving a terminology from the UMLS, in which the concepts will appear without listing their source terminologies.

A main result of our study is that the UMLS is used more as a terminology, an unintended use, than as a mapping mechanism between sources, an intended use [3], by a ratio of 7:4. Universities lead in using the UMLS as a mapping tool (47%). Healthcare employees are most likely to use the UMLS as a terminology (90%).

C Future Agenda

We found that 49 out of the 53 users using the UMLS as a terminology, and two others, want the NLM to derive a terminology from the UMLS. Healthcare and university employees are most interested and Software companies are least interested in the derived terminology. Genomics is leading in the request for expanded coverage, showing users' preference for more genomic concepts. To achieve this, the NLM needs to integrate more genomic source terminologies into the UMLS, beyond the integration [30] of GO [31]. Most other requested subjects are already represented in the UMLS, except for mapping and globalization.

Wrong or missing semantic type assignments, hierarchical relationships and associative relationships concern users more than other errors. Missing knowledge elements concern users less than modeling errors.

Users show substantial interest in the Semantic Navigator and in the need for an interface indicating multiple parents. It would definitely be helpful if the META interface were improved to support such a feature in the indented hierarchy display. Let us further illustrate the benefit of indicating ancestors with multiple parents in the indented list showing ancestors in the META interface. The existing UMLSKS META web interface suffers from overwhelming redundancy. First, the parents and ancestors are listed for each UMLS source terminology, causing a lot of repetition. Even though it makes sense to describe the hierarchy for each UMLS source, for a user who does not care about the sources, the repetition is overwhelming. Second, even for each source separately, each different ancestral path is listed and parents are listed multiple times in the parent list according to their appearances in the ancestral paths. To illustrate this, we show the data for just one source terminology the NCI for the concept Ectoderm. Figure 5 show the ancestors of Ectoderm, limited to NCI, using the style of the Semantic Navigator interface. As we see, Ectoderm has two parents Trilaminar Embryonic Disc and Embryonic Tissue. Each of these parents in turn has two parents. Furthermore,

Embryonic Disc, a grandparent has also two parents. Altogether there are 5 ancestral paths from Ectoderm to Embryonic Structure or System in Figure 5. Each of these paths is fully displayed in the UMLSKS interface. Furthermore these repetitions appear when listing the 5 parents. Moreover the same 4 children of Ectoderm are listed 5 times according to the 5(!) paths. Note that all this information is given even though the user asked only about Ectoderm.

Figure 5: The (NCI) Ancestors of Ectoderm Shown according to the Semantic Navigator Style

Embryonic Structure or System Embryonic Structure Embryo *Embryonic Tissue Ectoderm	Embryonic Structure or System Other Embryonic Structure *Embryonic Disc *Trilaminar Embryonic Disc Ectoderm

Figure 6 Indented ancestral paths with * indicating multiple parents

According to our proposal, only two-ancestral paths would be listed, one per parent (see Figure 6).

In our view, this limited indented list of ancestors in which ancestors with multiple parents are indicated with a (*) will suffice for most users interested in the concept Ectoderm. Those who need more information will, for example, be able to click on the starred parent Trilaminar Embryonic Disc and find that its other parent is Gastrulla, which was missing in the two ancestral paths of Figure 6. Such a compact display of ancestors will be more effective for users. It will still enable users to obtain further information in a way which directly points to the missing information, e. g. finding that Gastrulla is a grandparent of Ectoderm, while currently the details are buried in the repetitive lists of ancestors. This example also demonstrates the power of the Semantic Navigator as a graphical interface, capturing the hierarchical environment of a concept.

Concerning goals for the future of the UMLS, auditing was the most important task, followed by design of a derived terminology. Interestingly, expansion of coverage, where most of a putative budget is spent by NLM, is just third. Those results suggest that NLM should reconsider the priorities concerning the UMLS project.

D Limitations

The study's main limitation is based on the channel we used to disseminate the questionnaire. The channel through which we distributed the questionnaire could be said to introduce a bias into the results because the mailing list included about 600 members, while there were UMLS users who didn't join the mailing list and didn't receive our questionnaire at all. Those users might provide some valuable data that we are lacking in this study. Furthermore, the distribution of the members of the mailing list, e.g. according to various demographic and employment options, does not necessarily represent accurately the whole group of

UMLS users. However, the mailing list is the major channel the NLM uses to contact the UMLS users. For example, new UMLS releases and problems with the UMLS server were regularly and solely announced via the UMLS user list.[3] Thus, this UMLS mailing list was a natural and effective way to recruit UMLS users for our study, especially, since the study was neither initiated nor supported by NLM, to assure its independence.

Another limitation of our study is that from the 600 users of the mailing list, the number of the respondents was limited to approximately 12%. The low percentage may indicate unwillingness or lack of time for filling in a questionnaire estimated to require 25-30 minutes. For comparison, during the first 11 months of 2004, only 128 (about 21%) members, excluding NLM staff, were involved in the discussions transacted by email.

To get a perspective regarding the response rate obtained in our study, we looked for literature regarding email survey response rates. In a classical paper by Kim Sheehan [33], the dependency of the response rate on 5 aspects was studied. The aspects relevant to our study are: the year in which the study was undertaken, the number of questions and the number of follow-up contacts. Her study reviews 31 surveys conducted during 1986-2000, 25 of which were conducted during 1995-2000. It shows a clear trend of decline, in response rate, over the years. Sheehan attributes this decline to the decrease in novelty of email surveys with the spread of internet use, the increase in solicitation email in general and over surveying in particular. The author predicts that the declining trend will continue. The review reports 46%, 31% and 24% responses rate during the periods, 1995/6, 1998/9 and 2000. The question is what decline can be expected for 2004 when our study was conducted.

An approximation by linear regression of the 1995-2000 results of [33] suggests 10% for 2004. This estimate seems to be supported by recent publications. For example, in [34], it is reported that only 9% of the original study sample responded. In a commercial marketing website of Beeliner Survey [35], an

average response rate for electronic surveys is reported as 10%~20%. Our study response rate of 12% fits within these results.

In [33] the number of questions was found to be the second strongest predictor for the response rate. In our study there were many questions (26) and furthermore, the questions in the agenda part involved long explanations (see Appendix II). It is likely that the length of the questionnaire caused the reduced response rate. However, we wanted to receive responses for these complex questions, even at the price of a lower response rate, as they may give users' perspectives on the directions preferred for the UMLS project in the future. We sent a follow-up email to respondents and the number of responses was increased from 50 to 70. This phenomenon of positive influence of follow-up notification on the response rate is in line with the results of [33].

Due to the limitation of the number of returned questionnaires, the respondents may not accurately represent the population of the mailing list. A bias may exist in that there may have been disproportional numbers from certain kinds of organizations. For example, there were no respondents from any pharmaceutical companies. Employees at pharmaceutical companies might be working under stricter rules concerning activities such as filling in a voluntary questionnaire during working hours. However, we looked at the email addresses of users who posted emails to the UMLS mailing list during 2004. Many of those emails were from work addresses. Only one such user had an email address indicating a pharmaceutical company subsidiary (in Europe). This anecdotal information is in line with our finding that UMLS users from the pharmaceutical industry did not respond to our questionnaire. Maybe it indicates a low number of UMLS users from the pharmaceutical industry.

A bias might also existed with regards to the questions about age and education level distribution. This may be due to a typical situation in a laboratory or research group where only the senior people may hold a UMLS license and subscribed to the UMLS mailing list. Other less senior UMLS users in such a group may use the same

licenses. It is possible that such users were underrepresented in the study population.

Only a few responses, some of which were irrelevant, were obtained to the free text question about desired improvements in the UMLS interface. Maybe wording the question with "UMLS graphical user interface" would have been clearer.

V. CONCLUSIONS

In the final analysis, the presented study confirms that the UMLS is used to access its source terminologies and to map among them, which are two intended uses of the UMLS. However, we find that users access mainly just a few popular source terminologies and are slow to adapt transparent access via the rich release format. Users report many purposes of use. The leading categories are Terminology Research, Information Retrieval, Terminology Translation, UMLS Research and Natural Language Processing. The survey also shows that auditing the correctness of the UMLS and the design of a UMLS-based terminology are primary concerns of the existing user base. The latter is expected, since three quarters of the users actually use the UMLS as a terminology, even though it is not one. With regards to UMLS interfaces, may users agree with the suggestion that an indented hierarchy should mark concepts with multiple parents.

ACKNOWLEDGMENT

Carolyn Tilley of NLM UMLS Support provided helpful feedback for an early version of the questionnaire. However, this study was neither initiated nor supported by the NLM to assure its independence.

Dr. George Hripsack of Columbia University provided valuable advice regarding the methodology of the study and the paper.

REFERENCES

[1] Bodenreider O. The Unified Medical Language System (UMLS): integrating Biomedical terminology. Necleic Acids Res 2004; (32):D267-70. (Database issue)

[2] Humphreys B, Lindberg DAB, Schoolman HM, and Barnett GO. The Unified Medical Language System: An informatics research collaboration. JAMIA, 5(1), pp.1-13, 1998.

[3] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Meth Inform Med 1993;32(1):281—91.

[4] umls-users@umlsinfo.nlm.nih.gov (it has been changed to UMLSUSERS-L@list.nih.gov from 9/1/2005)

[5] Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc 1993;81(2):217—22.

[6] Tuttle MS, Sherertz DD, Olson NE, Erlbaum MS, Sperzel WD, Fuller LF, et al. Using META-1, the first version of the UMLS Metathesaurus. In: Proceedings of the 14th Annual SCAMC, 1990, p. 131—5.

[7] McCray AT. Representing biomedical knowledge in the UMLS semantic network. In: Broering NC, editor. High-performance Medical Libraries: advances in information management for the virtual era. Mekler; 1993. p. 45—55.

[8] http://www.nlm.nih.gov/research/umls/

[9]Noy NF and Carole D. Hafner CD. The State of the Art in Ontology Design: A Survey and Comparative Review. AI Magazine 18(3): Fall 1997, 53-74.

[10] McGuinness D, Ontologies Come of Age. In Fensel D, Hendler J, Lieberman H, and Wahlster W, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. Pp.171-197, MIT Press, Cambridge, MA, 2002.

[11] McCray AT and Miller RA (eds.). Making the conceptual connections: The UMLS after a decade of research and development. Special Issue of the Journal of the American Medical Informatics Association, 5(1), January/February 1998.

[12]Unified Medical Language System Bibliography. http://umlsinfo.nlm.nih.gov/bibliography.html. US Dept. of Health and Human Services, NIH, National Library of Medicine.

[13] Perl Y and Geller J, Guest (eds) Research on Structural Issues of the UMLS. Special issue of the Journal of Biomedical Informatics, 36(6), Dec. 2003.

[14] Hollis KF, The Unified Medical Language System from the National Library of Medicine: A review of practical uses for a biomedical information access system. http://leep.lis.uiuc.edu/publish/kfultz/451LE/ UMLSproject.htm.

[15] White Paper: UMLS Metathesaurus Rich Release Format. http://www.nlm.nih.gov/research/umls/white_paper.html

[16] Hole WT, Carlsen BA, Tuttle MS, Srinivasan S, Lipow SS, Olson NE, Sherertz DD, Humphreys BL Achieving "source transparency" in the UMLS Metathesaurus. Medinfo. 2004;11(Pt 1):371-5.)

[17] Perl Y and Geller J, Guest Editor's Introduction to the Special Issue: Research on Structural Issues of the UMLS—Past, Present, Future, Journal of Biomedical Informatics, 36(6) pp. 409—413, Dec. 2003.

[18] Cimino JJ, Battling Scylla and Charybdis. The search for redundancy and ambiguity in the 2001 UMLS Metathesaurus. In: S Sakken, editor. Proc 2001 AMIA Annual Symposium. Washington, DC; 2001. p.120-124.

[19] Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. In: Bakken S, editor. Proc 2001 AMIA Annual Symposium. Wachington, DC; 2001. p.57-61.

[20] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. Journal of Biomedical Informatics 2003;36(6):450-461.

[21] Gu H, Perl Y, Geller J, Halper M, Liu L, Cimino JJ. Representing the UMLS as an OODB: Modeling Issues and Advantages. JAMIA 2000; 7(1):66-80. Selected for repint in R. Haux, C. Kulikowski, eds.: Yearbook of Medical Informatics, International Medical Informatics Association, Rotterdam, 2001:271-285.

[22] Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artificial Intelligence in Medicine 2004; 31(1) :29-44

[23] Peng Y, Halper M, Perl Y, Geller J. Auditing the UMLS for redundant classification: In: Proc. 2002 AMIA Annual Symposium. San Antonio, TX; 2002. p.612-616.

[24]McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med. 1995 Mar;34(1-2):193-201.

[25] Geller J, Gu H, Perl Y, Halper M. Semantic refinement and error correction in large terminological knowledge bases, Data & Knowledge Engineering, 45(1), 2003, pp. 1-32.

[26] Min H, Perl Y, Chen Y, Hapler M, Geller J, Wang Y. Auditing as Part of the Terminology Design Life Cycle, JAMIA. 2006; 13:676-690

[27] Tang D, Stolte C, Bosch R. Design Choices when Architecting Visualizations, Information Visualization 5(2) 65-79, 2004

[28] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. In: Proceedings of Medinfo 2001, London, UK, September 2001, p. 171—5.

[29] Bodenreider O. Personal communication.

[30] Lomax J. and McCray A.T. Mapping the Gene Ontology into the Unified Medical Language System. Comparative and Functional Genomics, 5(5):345-361, 2004

[31] G.O. Consortium. Creating the Gene Ontology resource: Design and implementation. Genome Res., 11:1424-1433, 2001.

[32] umls-announces-l@list.nih.gov

[33] Sheehan K E-mail Survey Response Rates: A Review Journal of Computer Mediated Communication 6 (2), 2001

[34] Norman, A. T., and Russell, C. A.. The pass-along effect: Investigating word-of-mouth effects on online survey procedures. Journal of Computer-Mediated Communication, 11(4), article 10, 2006

[35] Beeliner Survey Software http://www.beelinersurveys.com/newsroom/mediafaqs.html (Oct. 5, 2006)

APPENDIX I

Table A1: Age Group

Age	Number (%)
51~60 yrs. Old	21 (31%)
41~50 yrs. Old	17 (24%)
31~40 yrs. Old	15 (21%)
30 yrs. Old or less	9 (13%)
Above 60 yrs. Old	4 (6%)
Unknown	3 (4%)

Table A2: Organization Size

Size	Number (%)
1001~5000	20 (29%)
More than 5000	18 (26%)
501~1000	9 (13%)
11~50	8 (11%)
10 or less	5 (7%)
51~100	5 (7%)
101~500	3 (4%)
No answer	2 (3%)

Table A3: The Distribution of User Numbers according to Number of Areas Selected (See Figure 1)

# of areas	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# of users	4	16	4	6	4	5	2	6	8	1	2	2	1	1	1	7

APPENDIX II

Questionnaire Part about UMLS Users, Use and Future Agenda (Part III)

Agenda:

- ***The UMLS is not a terminology but a terminological knowledgebase integrating many source medical terminologies. Therefore, it contains inconsistent knowledge from different source terminologies, which cannot be removed.***

1. Are you using the UMLS to map (i.e. translate) between its source terminologies, e.g. ICD10 & SNOMED CT?
   __Yes __No

2. Are you using the UMLS as a terminology?
   - ***For example, one may be looking up a user's term in the UMLS to obtain a list of concepts with matching names. (For example, if a user types in "pneumonia", and finds a list of 140 UMLS concepts).***

   __Yes __No

3. Would you be interested in the NLM designing a terminology derived from the UMLS? The remark below describes such a potential derived terminology.
   - In this derived terminology the information about the occurrence of terms in source terminologies will be removed. Thus, inconsistencies found in the UMLS could be removed from the derived terminology. Note that information regarding occurrence in source terminologies will still be available to a user in the UMLS, which will not change.

   __Yes __No

4. In which subject areas would you desire that the NLM expand the UMLS coverage?
   a) _____

b) _____
c) _____

- ***Auditing is the activity of ensuring the quality of the knowledge and correcting errors.***

5. How much do the following kinds of modeling errors bother you when encountered?
   - ***Examples are from the 2004AA release and are mainly from concepts assigned the semantic type* "Environmental Effects of Humans"*. Some examples may be debatable, but they are only used for demonstration purposes. Note that the role of auditing is just to raise questions. Only NLM has the authority to decide how to resolve problems found.***

a) Concept redundancy (two META concepts with the same meaning. e.g. "*Pollution (of environment)*" and "*Environmental pollution*")
Bothers me: __not at all; __a little; __moderately; __a lot.

b) Concept polysemy (one META concept with multiple meanings, e.g. "*Video recording*" is the "**Human-caused Phenomenon or Process"** of recording, but "*Video recording "* is also the recorded tape, which is a "**Manufactured Object"**)
Bothers me: __not at all; __a little; __moderately; __a lot.

c) Wrong hierarchical relationships (*a* IS-A *b,* where *a* is not a correct specialization of *b*, e.g., "smog" IS-A "*Social problem, NOS*" (removed in 2004AB release)).
Bothers me: __not at all; __a little; __moderately; __a lot.

d) Wrong associative relationships (wrong relationship name or wrong target concept, e.g. META concept "*Specimen from stomach obtained by incisional biopsy*" with a relationship "**specimen source topography,**" has wrong target "*Large intestinal structure,*" instead of "*Stomach*".)
Bothers me: __not at all; __a little; __moderately; __a lot.

e) Wrong semantic type assignments (E.g. "*College environment*" is assigned "**Manufactured Object" and "Organization,"** both wrong.)
Bothers me: __not at all; __a little; __moderately; __a lot.

f) Redundant semantic type assignment (Parent and child semantic type are both assigned to a concept. E.g. "*Carboxybenzyl-leucyl-leucyl-leucine vinyl sulfone*" was assigned "**Organic Chemical;**" "**Amino Acid, Peptide, or Protein**" and "**Pharmacologic Substance.**" The semantic type "**Organic Chemical**" is a parent of "**Amino Acid**, **Peptide, or Protein.**" Thus, the assignment "**Organic Chemical**" should be removed by UMLS rules.)
Bothers me: ___ not at all; ___ a little; ___ moderately; ___ a lot.

6. How much do the following kinds of missing knowledge bother you when encountered?
   - ***Usually, a terminological knowledge element such as concept, definition, relationship is missing in the UMLS since it is not provided by any of its source terminologies. Our question is whether you are bothered by the fact of such an element not appearing in the UMLS (independent of the reason), when by context of other existing elements, you, as a user, expect it.***

a) Missing concepts (expected in the context of existing concepts, e.g. there exist "*cigar,*" "*cigarette*" and "*Second hand cigarette smoke.*" But "*Second hand cigar smoke*" does not exist in the META.)
Bothers me: __not at all; __a little; __moderately; __a lot.

b) Missing definitions (e.g. "*Environmental pollution*" doesn't have a definition)
Bothers me: __not at all; __a little; __moderately; __a lot.

c) Missing synonyms (e.g. "*Environmental pollution*" is missing a synonym "*Pollution (of environment)"*)
Bothers me: __not at all; __a little; __moderately; __a lot.

d) Missing hierarchical relationships (e.g. "*Industrial smog*" should be a child of "*Smog*")
Bothers me: ___ not at all; ___ a little; ___ moderately; ___ a lot.

e) Missing associative relationships (e.g. "*Industrial smog*" mapped from "*Air Pollution*" (same as for "*Smog*")).
Bothers me: __not at all; __a little; __moderately; __a lot.

f) Missing semantic type assignments (e.g. "*Smog*" is not assigned to "**Environmental Effects of Humans,**" even though its children and parents are.)
Bothers me: ___ not at all; ___ a little; ___ moderately; ___ a lot.

- **The UMLSKS META Interface *displays an indented hierarchy of the ancestors of a concept.* The NLM Semantic Navigator *provides a diagrammatic display of a concept and its related concepts.***

7. Currently, multiple parents are displayed by the UMLSKS META interface only for the focus concept but not for its ancestors in the indented hierarchy. Would you want the NLM to enhance the META interface to mark a concept with multiple parents in the indented hierarchy (similar to the way MS Windows marks a concept with children by a "+")?

- **This way, when a user sees a concept in the indented hierarchy marked with multiple parents, he/she may switch to the NLM's Semantic Navigator, where multiple parents are displayed for ancestors as well, or view that ancestor as the focus concept.**
__Yes __No

8. Would you like to see the NLM's Semantic Navigator for the UMLS 2004 version? **(It is available for 1998 ~ 2003**)
__Yes __No

9. What other enhancements of the META interface would you suggest?
_____
_____

10. Assuming a given budget for future UMLS work beyond maintenance, please specify the percentages that would you like the NLM to allocate to:
__Expansion of coverage of new subjects
__Design of a derived terminology from UMLS (see question No.19)

__Auditing for the improvement of the quality of the knowledge
__Design of better interfaces
__Others (please specify)_____

---

[1] Responses were written in free text.

[2] The Semantic Navigator is available every year only for the first (AA) release of the UMLS and is made available with a delay while later releases of the UMLS are already available [29]. At the time when we sent out the survey, only versions of the Semantic Navigator from 1998-2003 were available.

[3] As of August 2005, the NLM maintains two separate UMLS mailing lists, one for official announcements [32] and the other for announcements and discussions. When we conducted our study, there was just one unified list.