Who is Using the UMLS and How – Insights from the UMLS User Annual Reports

Kin Wah Fung MD MS MA, William T. Hole MD, Suresh Srinivasan MS

National Library of Medicine, Bethesda, Maryland

kwfung@nlm.nih.gov

The NLM's UMLS resources are available to users free of charge under a license that requires submission of an annual report on their usage. A new web-based template was used to collect users' annual reports for the calendar year 2004. Out of 2,677 licensees, 1,427 (53%) submitted their annual reports through the web template. This represented a five-fold increase in the reports submitted compared to previous years. The information collected via the web template was more structured, more complete and easier to analyze. The main results from the 2004 annual reports are summarized and discussed. They are being used to guide UMLS developments.

INTRODUCTION

This year marks the 20th anniversary of the National Library of Medicine's (NLM) Unified Medical Language System (UMLS) project. Started in 1986, the purpose of the UMLS project is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of sources. [1-5] One of the barriers that need to be overcome is the multiple ways in which the same meanings are expressed in different information sources and by the users themselves. The solution offered by the UMLS is a Metathesaurus that inter-connects a myriad of biomedical vocabularies. The 2006AA release of the UMLS contains over 1.2 million concepts with 6 million names from more than 100 source vocabularies in 17 languages. In addition to the Metathesaurus, the other two knowledge sources that make up the UMLS are the Semantic Network and the SPECIALIST lexicon/lexical tools.

NLM has always been interested in gathering usage information and feedback from direct users of the UMLS. One source of such information is the published literature. [6] A quick estimation of the number of UMLS-related publications can be obtained by a PubMed/Medline search using "Unified Medical Language System" and "UMLS" as the search terms. As shown in Figure 1, there has been a steady increase in the number of UMLS-related publications over the years.

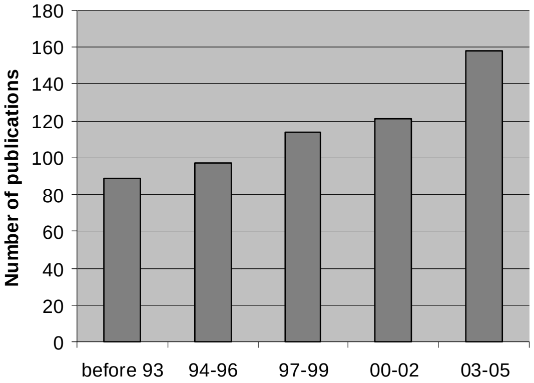

Figure 1: Steady increase in UMLS-related publications in PubMed/Medline

Another important source of information is the annual report submitted by UMLS users. The UMLS is provided free of charge by NLM but users need to obtain a license. One requirement of the license is to submit an annual report on the usefulness of the UMLS. Annual reports prior to 2004 were submitted via email and were largely unstructured. Analysis of the reports was time-consuming. Typically, less than 10% of licensees submitted their reports every year. From 2005 onwards, an on-line template was used to collect annual reports. This paper summarizes the main results from the first web-based UMLS user annual reports.

METHODS

The annual report template was created and tested by internal NLM users. The use of the web form to collect annual reports was announced through the UMLS listserv and emails to individual UMLS licensees. The web template was released to the public in February 2005. Subsequent reminder emails were sent to non-responders in March and July. Users were notified that they had to submit their reports to keep their licenses active. The collection process ended in August.

Users of the template were required to provide their license numbers for identification. The annual report was divided into five sections. All users were required to complete the first section (main report) which focused on three areas: user profile (e.g. affiliation, computing environment), usage of the UMLS and feedback on user support and communication. The other four sections were each directed to a specific UMLS component, namely the Metathesaurus, Semantic Network, SPECIALIST lexicon/lexical tools and UMLS Knowledge Source Server (UMLSKS). These four sections were optional and users were encouraged to answer the questions about the specific resources they had used.

RESULTS

There were altogether 2,677 registered UMLS licensees up to January 2005. Among them, 1,427 (53%) submitted their annual reports through the web template. The following is a summary of the data from the main report section and the section on the Metathesaurus. Percentages were calculated based on the total number of licensees responding to a particular question. When multiple answers were allowed for a question, the sum of the percentages of that question would be more than 100%.

Geographic distribution

Information on geographic location was collected previously when users applied for the UMLS license. Table 1 shows the geographic location of all UMLS licensees and those who have submitted their 2004 annual reports.

Geographic region	All UMLS licensees (% of total)	Licensees with annual report (%)
North America - U.S.A. - Canada	 1,943 (73%) 78 (3%)	 1,093 (74%) 36 (3%)
Europe	368 (14%)	210 (15%)
Asia and Australasia	231 (9%)	105 (7%)
Central and South America	45 (2%)	16 (1%)
Africa	12 (0.4%)	3 (0.2%)
Total	2,677 (100%)	1,427 (100%)

Table 1. Geographic distribution of UMLS licensees

Licensee affiliation and main activity

Most of the users were affiliated with academic (40%) or commercial (26%) entities. (Figure 2) The main activities of the affiliated entities were mostly research (26%), software development (20%), health care provision (17%) and education (16%). (Figure 3)

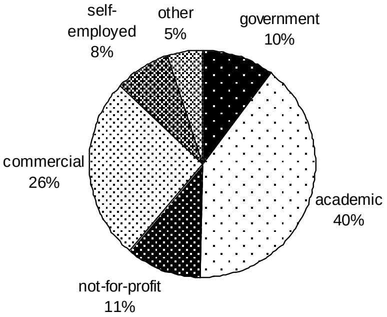

Figure 2. Affiliation of UMLS licensees

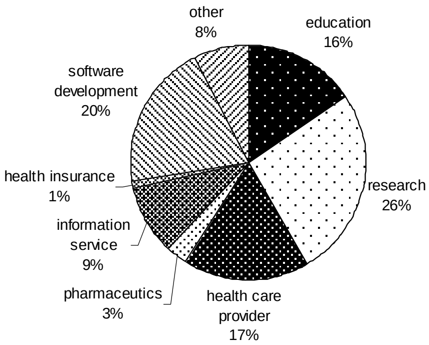

Figure 3. Main activity of affiliated organization

Computing environment

Windows was the most common operating system (89%), followed by Linux (26%), UNIX (18%), Macintosh (10%) and others (2%). As for programming languages, Java was the commonest (55%), followed by C++ (25%), Perl (25%), Visual Basic (24%) and others (30%). Database management systems included MySQL (35%), Microsoft SQL (33%), Oracle (31%), Microsoft Access (31%) and others (20%).

Usage of the UMLS

A total of 675 licensees (47%) had not started using the UMLS. Among the 752 (53%) who had, the median duration of usage was 12 months with a range of 1 – 180 months (mode 12 months, mean 31 months, skewed distribution to the left, figure 4). The apparent spikes in the figure are most likely the result of users rounding off their duration of usage to the year.

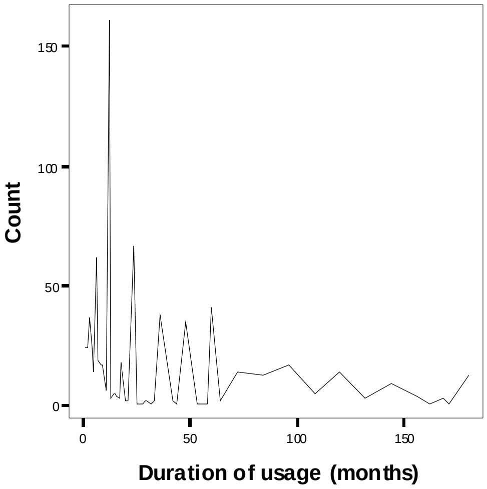

Figure 4. Duration of UMLS usage

Most users used the UMLS for the processing of clinical information, the commonest categories being findings and diagnosis (55%), procedures (34%), laboratory tests (29%) and drugs (29%). As for non-clinical information, genomics/proteomics (15%) and bibliographic information (14%) were the commonest. (Table 2)

Type of information		% of users
clinical	findings/diagnosis	55%
	procedures	34%
	laboratory tests	29%
	drugs	29%
	clinical guideline	21%
	clinical outcome	18%
	nursing info	10%
	other	11%
non-clinical	genomics/proteomics	15%
	bibliographic info	14%
	consumer health	10%
	billing	7%
	other	9%

Table 2. Types of Information that the UMLS was used to process

The UMLS was most frequently used for terminology research (53%), mapping between terminologies (35%) and creation of a local terminology (33%). Other uses included: information indexing/retrieval (31%) and natural language processing (21%). (Table 3) Among the three knowledge sources in the UMLS, the Metathesaurus was most commonly used (94%), followed by the Semantic Network (41%) and the SPECIALIST lexicon/lexical tools (28%). All three were used by 19% of users.

Area of usage	% of users
terminology research	53%
terminology mapping	35%
terminology building	33%
information indexing/retrieval	31%
natural language processing	21%
knowledge acquisition	20%
concept discovery	19%
terminology service	14%
access to terminologies	13%
terminology publishing	6%
other	5%

Table 3. Area of usage of the UMLS

User support and communication

The written documentation of the UMLS was considered adequate by 85% and inadequate by 15% of users. Apart from the written documentation, NLM also provides user support through individual email, answering questions posted on a listserv and telephone enquiry. Altogether 226 (16%) users had experience with one or more of these user support services. Among the email users, over 90% rated the support service as either very useful or sometimes useful. Ratings for the other support services were similar. (Table 4)

User support service	Number of users	Rating of support service			
		Very useful	Some-times useful	Rarely useful	Not use-ful
Email	179	103	60	6	10
Listserv	64	31	28	5	0
Phone	54	41	8	2	3
Others	10	8	2	0	0

Table 4. Rating of user support services

Feedback on the Metathesaurus

Among all users who submitted their annual reports, 328 (23%) also completed the section on the Metathesaurus.

Access

The majority of users accessed the Metathesaurus either by local installation or browsing via the UMLSKS website (local installation only 38%, UMLSKS browsing only 40%, both 10%). 12% of users used the Application Program Interface (API) provided by the UMLSKS, either alone or in combination with local installation and/or browsing.

Local installation

As for the file format used in local installation, almost the same number of users used the old Original Release Format (ORF) as the new Rich Release Format (RRF) (ORF only 48%, RRF only 46%, both 6%). To see whether the choice of file format was related to the duration of UMLS usage, we compared the durations of UMLS usage of the two groups. The mean durations of UMLS usage were 39.5 and 38.8 months for ORF and RRF users respectively. The difference was not statistically significant (p=0.99 Mann-Whitney two-tailed non-parametric test). MetamorphoSys, the installing and subsetting tool of the Metathesaurus, was considered to be adequate for its purpose by 81% of users.

Vocabularies and contents usage

47% of users used all vocabularies in the Metathesaurus. 44% of users used SNOMED CT and level 0 sources (vocabularies with no additional restriction), either alone or in combination (SNOMED CT only 24%, level 0 sources only 12%, both 8%). Apart from SNOMED CT, the most frequently used vocabularies were: RxNORM, CPT, ICD9/10, MeSH and LOINC. As for the coverage of the Metathesaurus, 77% of users thought that it was adequate. The areas most frequently thought to be inadequately covered were: genomics/proteomics, public health and consumer health. Non-English contents of the Metathesaurus were used by 17% of users. Among them, the most commonly used foreign language sources were German, French and Spanish. The usage pattern of the various types of Metathesaurus contents is summarized in Table 5.

Type of Metathesaurus contents	**% of users**
Concept names and synonyms	93%
Hierarchical relationships	70%
Non-hierarchical relationships	36%
Unique identifiers	56%
Definitions	56%
Mappings	56%
Co-occurrence data	17%

Table 5. Usage of the various types of Metathesaurus contents

DISCUSSION

UMLS user annual reports are an important source of information that provides guidance in future developments. The way in which these reports were collected previously (via email) was unsatisfactory because of the low response rate and difficulty in analysis. The new web-based template has resulted in significant improvements. Firstly, there was a five-fold increase in response rate. This was probably due to two factors. The use of a database to track responses allowed us to send timely reminder emails to non-responders. Only 30% of the annual reports were received before the first reminder was sent out. Moreover, the fact that filling in a web form is easier than composing a free-form email probably also contributed to the higher response rate. The second advantage of the web-based collection of annual reports was better data quality. The structured nature of the reports ensures that important areas are covered. A uniform data structure facilitates collective analysis will allow meaningful comparison of results in subsequent years. These advantages of a structured, web-based questionnaire for collection of user feedback are certainly not limited to the specific use case depicted here. Any organization delivering a product or service to users should be able to benefit similarly.

The fact that only half of all licensees submitted annual reports is no surprise. Since the UMLS license is free-of-charge and easily obtainable on-line, some licensees may only find out after obtaining the license that the UMLS is not suitable for their purpose or that they do not have time or technical knowledge to explore this resource further. This group of licensees does not actually use the UMLS, and mostly will not submit annual reports. After the deadline of annual report submission in August 2005, the licenses of over 1,000 non-responders were inactivated. Among them, only seven subsequently requested their licenses to be reactivated. This shows that most of the non-responders are no longer using the UMLS. In other words, those who have submitted annual reports are the active UMLS users. The collection of annual reports turned out to be an effective way of cleansing our licensee database of inactive users.

Despite the long history of the UMLS project, half of the users have only been using it for 12 months or less. One possible explanation is that some users only need to use the UMLS for a short period of time (e.g. students using the UMLS for a project). Another factor that had contributed to this usage pattern was the surge in the number of new licensees in 2004. This was probably related to the first inclusion of SNOMED CT in the UMLS in early 2004, as a result of an agreement between the U.S. Department of Health and Human Services and the College of American Pathologists to make SNOMED CT available to U.S. users at no cost. [7] The observation that SNOMED CT was used by a large proportion (79%) of UMLS users seems to support this correlation. It will be interesting to see how the duration of usage pattern might change in subsequent years.

Far more users used UMLS to process clinical information (81%) than non-clinical information (45%). This is probably related to the composition of the Metathesaurus, which consists mainly of clinical vocabularies. Non-clinical subject areas (e.g. genomics/proteomics and consumer health information) were most frequently named as examples of inadequate coverage. It seems that there is a demand to expand the coverage of UMLS in these areas. However, the availability of high quality vocabularies and their willingness to be included in the UMLS will determine to what extent the UMLS can satisfy this need.

Concerning user support, over 90% of users rated our user support services as helpful. However, the written documentation was considered unsatisfactory by 15% of users. A common suggestion to improve the documentation was to include a “getting started” tutorial, or to make it less technical and less complex. These suggestions are understandable as the UMLS is indeed a complicated resource and the learning curve can be quite steep. In view of this, NLM offers regular classes to help users understand and use the UMLS. Some of the materials used in these classes will be made available on-line in the near future. Interestingly, there were also some appeals for more detailed and more technical documentation. This reflects the diversity of the users in terms of their level of knowledge and technical capability. It is an ongoing challenge to produce documentation that satisfies the needs of most users.

The fact that many users were still installing the Metathesaurus in the old ORF format was a bit surprising. Since the 2004AA release, RRF has been the default output format of MetamorphoSys. RRF has significant advantages over ORF, among which is the ability to represent information in source vocabularies more completely and accurately (source transparency). [8] Users are encouraged to use RRF but ORF will be supported for backward compatibility. One would normally think that ORF users were more likely to be those who had started using the UMLS before RRF was introduced. However, it turned out that there was no significant difference between the durations of UMLS usage of the RRF and ORF users. In addition, among the users who had used UMLS for 12 months or less (i.e. RRF was already available to them when they started using the UMLS), there were almost as many ORF users as RRF users. One possible explanation of why new users chose to use ORF is that they were influenced by other users who were already using ORF. It will be interesting to see how the proportion RRF users might change in future.

Finally, it is worth mentioning that UMLS licensees only represent direct users of the UMLS resources. Their number does not truly reflect the total number of UMLS end users. One licensee may develop applications or terminology services used by thousand of UMLS end users.

CONCLUSION

The new web-based template proved to be a better means of collecting annual reports from UMLS licensees. The response rate was much improved and so was the quality of the data obtained. NLM will continue to use this method to collect feedback from UMLS users.

Acknowledgements

The authors thank Carolyn Tilley and Rachel Kleinsorge for their contributions in the design, implementation and administration of the UMLS licensee annual report web template.

References

1. Lindberg DA, Humphreys BL. Toward a unified medical language. Proceedings of the 7th International Congress of Medical Informatics Europe '87 1987:23-31.

2. Humphreys BL, Lindberg DA. Building the Unified Medical Language System. Proc Annu Symp Comput Appl Med Care 1989:475-80.

3. Humphreys BL, Lindberg DA, Hole WT. Assessing and enhancing the value of the UMLS Knowledge Sources. Proceedings - the Annual Symposium on Computer Applications in Medical Care 1991:78-82.

4. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of Information in Medicine 1993;32:281-91.

5. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 1998;5:1-11.

6. Current Bibliographies in Medicine 96-8: Unified Medical Language System. Available from: http://www.nlm.nih.gov/archive/20040831/pubs/cbm/umlscbm.html.

7. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inform Assoc 2005;12:486-94.

8. 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;11:371-5.