Ubiquity: Technologies for Better Health in Aging Societies
A. Hasman et al. (Eds.)
IOS Press, 2006

Specific classification of *e*library resources says more about users' preferences

Judas Robinson[a1], Simon de Lusignan[a], Patty Kostkova[b], Bruce Madge[c], Lesley Southgate[a]
[a] *St. George's, University of London, Cranmer Terrace, London, SW17 0RE*
[b] *City eHealth Research Ctr., Inst. of Health Science, City University, London, EC1V 0HB*
[c] *British Medical Association Library, BMA House, Tavistock Square, London WC1H 9JP*

Abstract: Background: Medical Subject Headings (MeSH) are a hierarchical taxonomy of over 42 000 descriptors designed to classify scientific literature; it is hierarchical with generic high order headings and specific low order headings. Over 1,000 resources in the Primary Care Electronic Library (PCEL – www.pcel.info) were classified with MeSH.
Methods: Each of the entries or resources in the primary care digital library was assigned up to five MeSH terms. We compared whether the most generic or specific MeSH term ascribed to each resource best predicted user preferences.
Results: over the four month period analysed statistically significant differences were found for resources according to specific key MeSH terms they were classified by. This result was not repeated for generic key MeSH terms. Conclusions: Analysis of the use of specific MeSH terms reveals user preferences that would have otherwise remained obscured. These preferences are not found if more generic MeSH terms are analysed.
Keywords: Medical Subject Headings; Libraries, Digital; Primary Health Care

1. Introduction

We chose Medical Subject Headings (MeSH) as the controlled vocabulary to index the Primary Care Electronic Library (PCEL) [1]. There are a range of alternatives we could have used: (1) A disease or clinical classification, e.g. International Classification of Diseases (ICD) [2] and the Systematized Nomenclature of Medicine (SNOMED) [3]; (2) A procedure coding system, e.g. Office of Population, Censuses and Surveys – Classification of Surgical Operations and Procedures – 4th Revision (OPCS-4); (3) Another library classification e.g. Dewey Decimal Classification hierarchy [4]; or (4) Used a metathesaurus e.g. Unified Medical Language System (UMLS); or (5) a combination of tools [5]. We selected MeSH because we wanted a system with a hierarchical structure which would enable those browsing the library to focus or broaden their search by browsing its hierarchy. There is some evidence that this orientates the user and can also reveal user choices [6]. , Using ICD-10 would provide a hierarchy [7] but has the disadvantage that it specifically classifies diseases; its clinical modification (e.g. ICD-10-CM) broadens its scope by including procedures yet does not have the breadth of MeSH or address areas like the type of publication. In addition there have been problems with discontinuation of codes between versions [8]. SNOMED CT or other clinical classification

[1] Judas Robinson, Primary Care Informatics, Division of Community Health Sciences, St George's - University of London, London SW17 0RE jsrobins@sgul.ac.uk

might offer advantages for linking to data within clinical records [9], but would also be difficult to link to scientific concepts or subjects. ULMS would meet our needs, however the wealth of concepts and linkages to clinical classifications within UMLS are probably too extensive for a single library resource [10].

The Primary Care Electronic Library (PCEL - www.pcel.info) contains over 1,000 abstracts of quality assured Internet resources for primary care. PCEL is the latest stage of development of digital library that has been online for 6 years [11]. Users can directly access resources or browse the MeSH tree for relevant material. Each resource has up to five MeSH subheadings or terms. These were either allocated by a qualified medical librarian or an academic supported by librarians at St. George's. We were unsure whether applying more generic (e.g. Cardiovascular disease) or specific MeSH subheadings (e.g. Myocardial infarction) was of more help to users in finding resources. We therefore decided to investigate whether generic or specific MeSH headings told us most about user preferences.

2. Method

Requests for resources were evaluated. Requests originating from identified users and search engines were separately analysed. Resource requests were compared with expected values; comparing whether the observed and expected usage was predicted by either the most generic or specific MeSH terms. Expected values were the numeric proportion of resources in each category of the MeSH tree top structure (this can be viewed and browsed from the NLM website [12]). Data was collected over a four month period. Log files were parsed and relevant data was stored in a database for analysis. We used the Chi-squared test to see whether the proportion of requested tests differed significantly from the expected proportion [13]; i.e. to indicate whether a group of resources was used disproportionately more or less than expected.

PCEL is written in ColdFusion and hosted on an Apache web server. Apache log files were used to collect data regarding site activity. In preparation for this evaluation the web site was altered so that all visits to pages relating to MeSH recorded relevant details in the log. The storage of MeSH terms was migrated to a hierarchical database architecture, using MeSH 2005 in ASCII format [14].

To test the face validity of the method results for an arbitrary taxonomy were also calculated. This was constructed by grouping resources according to their chronological entry in the database, the auto increment primary key of the table which contains resources. The results of the MeSH analysis, those of the arbitrary taxonomy, and resource requests were presented online.

3. Results

PCEL attracted 65998 requests in the four months under analysis. However, 45661 were from search engines. Approximately 31% of usage comes from identified users. As well as requests for resources, 775 requests were recorded for browsing the MeSH tree. The

number of times each resource was requested over the four month period was recorded. The full list of requests for resources is presented online [15]. Although some resources were consistently popular this was the exception rather that the rule. Resources which attracted more than two requests per month each month accounted for 37.2% of the total requests.

A comparison of the distribution of key MeSH terms (both specific and generic) and all MeSH terms used is presented online [15]. It shows that broadly speaking, the distribution of key MeSH terms is representative of all MeSH terms applied to resources. The results of Chi squared analysis for the key MeSH terms (both specific and generic) of requested resources are presented online[15]. The results for specific key MeSH terms are presented in table 2.

Five of the fifteen descriptors show consistent statistical differences over the period analysed: Biological Sciences, Health Care, Physical Sciences, Information Science and Anthropology, Education, Sociology and Social Phenomena. These five descriptors cover 66.5% of PCEL resources. Only three MeSH tree top descriptors identified by the most generic key MeSH terms show consistent statistical differences over the period analysed: Anatomy, Physical Sciences and Anthropology, Education, Sociology and Social Phenomena. These three descriptors only account for 12.8% of PCEL resources. The search engine requests showed no statistically significant differences from expected results.

The results of Chi squared analysis for an arbitrary classification of requested resources is presented online [15]. Of thirty results for categories over the time period only four show statistically significant differences from those expected. Analysing all categories for the month of April did not yield a statistically significant result. None of the categories of the arbitrary taxonomy show consistent statistical differences over the period analysed.

4. Discussion

Our results show that more specific MeSH terms applied to the resources of a digital library reveal user preferences. These user preferences are not defined by the analysis of more generic MeSH terms, and can serve to guide the indexing of material for the library. The findings also emphasise that it may be more useful to apply more specific MeSH terms to resources when classifying material. Classifying into taxonomies, such as MeSH, is a long established technique. Applying this idea to digital content is a practice which has recently grown in importance; adherents arguing that taxonomies complement traditional keyword searches and help users efficiently find data [16]. Although debate exists at to whether this encourages or inhibits the creative process of finding information, controlled vocabularies are a standard resource for information retrieval [17].

Over the period analysed the observed requests for two thirds of PCEL resources were statistically different to those expected on the basis of their most specific key MeSH terms: for the majority of resources the frequency of requests could be predicted on the basis of their MeSH classification. An arbitrary classification system showed no consistent statistical differences. This strengthens the case that the results for key specific MeSH

Table 2 - Differences in requests identified by MeSH tree top category

Tree Top MeSH Descriptor	Month.	Observed(O).	Expected(E).	0-E.	P
Anatomy	Jan	12	13	-1	0.7801
	Feb	22	13	9	0.0120 *
	Apr	15	13	2	0.5767
Organisms	Jan	7	14	-7	0.0596
	Feb	10	14	-4	0.2818
	Apr	12	14	-2	0.5904
Diseases	Jan	230	190	40	0.0013 **
	Feb	228	190	38	0.0023 **
	Apr	206	190	16	0.1996
Chemicals and Drugs	Jan	17	13	4	0.2642
	Feb	4	13	-9	0.0120 *
	Apr	5	13	-8	0.0255 *
Analytical, Diagnostic and Therapeutic Techniques and Equipment	Jan	22	45	-23	0.0004 ***
	Feb	45	45	0	1
	Apr	27	45	-18	0.0060 **
Psychiatry and Psychology	Jan	33	44	-11	0.0902
	Feb	40	44	-4	0.5378
	Apr	46	44	2	0.7580
Biological Sciences	Jan	243	162	81	0 ***
	Feb	229	162	67	1.0445e-8 ***
	Apr	230	162	68	6.2534e-9 ***
Physical Sciences	Jan	7	26	-19	0.0001 ***
	Feb	8	26	-18	0.0003 ***
	Apr	11	26	-15	0.0028 **
Anthropology, Education, Sociology and Social Phenomena	Jan	18	49	-31	0.0000 ***
	Feb	29	49	-20	0.0034 **
	Apr	12	49	-37	6.1742e-8 ***
Technology and Food and Beverages	Jan	2	8	-6	0.0332 *
	Feb	3	8	-5	0.0759
	Apr	1	8	-7	0.0129 *
Humanities	Jan	3	3	0	1
	Feb	4	3	1	0.5631
	Apr	0	3	-3	0.0828
Information Science	Jan	196	170	26	0.0294 *
	Feb	225	170	55	0.0000 ***
	Apr	206	170	36	0.0025 **
Persons	Jan	2	10	-8	0.0110 *
	Feb	3	10	-7	0.0261 *
	Apr	5	10	-5	0.1121
Health Care	Jan	259	292	-33	0.0230 *
	Feb	201	292	-91	0 ***
	Apr	248	292	-44	0.0024 **
Geographic Locations	Jan	0	12	-12	0.0004 ***
	Feb	0	12	-12	0.0004 ***
	Apr	27	12	15	0.0000 ***

N = 1051 * p<.05, ** p<.01, *** p<.001

terms are significant. The data for individual resource requests showed that it is the exception rather than the rule for resources to be requested more than two times over the four months analysed. This excludes regular requesting of resources as a factor in trends noticed for key MeSH terms. Also no comparison was found between the sample size for the MeSH tree top descriptor and the probability that observed results differed from expected. These results confirm that the MeSH taxonomy classifies Internet resources into functional groups, functional groups associated with user preferences.

Interestingly this user preference is to a large degree absent if the most generic MeSH terms are considered. Only 12.8% compared with 66.5% of PCEL resources were shown to have statistical differences associated with generic key MeSH terms compared with specific key MeSH terms. This would indicate that more specific entries in the MeSH hierarchy are better able to indicate user preferences. This perhaps should not come as a surprise as more specific MeSH terms are a more accurate indicator of the subject under consideration.

MeSH lends itself to hierarchical searching as well as analysis. Thus the relative frequencies of requests for resources identified by key MeSH terms can be classified into these 15 groupings for analysis. Such data provides information concerning the user preferences of PCEL users and can act as a guide for indexing material in the future. The results showed that PCEL users have preferences for resources described by the MeSH tree top descriptors 'Biological Sciences' and 'Information Science' and preferences against 'Physical Sciences', 'Anthropology, Education, Sociology and Social Phenomena', and 'Health Care'. On more specific levels of the MeSH tree, in January 2005, preferences were shown for 'Computing Methodologies', 'Medical Informatics', 'Health Occupations', 'Circulatory and Respiratory Physiology', 'Nursing', 'Digestive System Diseases', 'Respiratory Tract Diseases', and 'Cardiovascular Diseases'.

The results are limited by two factors. Firstly, the size of the MeSH taxonomy exceeds the number resources indexed in PCEL: there are over 42,000 terms in the MeSH vocabulary, and of these only 942 are used for PCEL's 1000 resources. The second limitation is the number of hits PCEL receives. This was sufficient for statistical analysis in January, February and April 2005, but fell short for the month of March.

We failed to identify other studies reporting on the differences between specific and generic MeSH terms applied to a browsable classification. The National Library of Medicine permits searching and browsing of MeSH classifications [18], but this functionality is not fully integrated with the hierarchical searching of MEDLINE. Another digital resource using the MeSH taxonomy is Organising Medical Networked Information (OMNI) [19]. Although the Resource Discovery Network (RDN), to which OMNI belongs, has published evaluation reports [20], it was difficult to find evaluation reports from the RDN hubs. Further studies are needed to see if user preferences are defined by more specific levels of classification using MeSH and other taxonomies.

5. Conclusions

Assigning MeSH terms to Primary Care Internet resources and allows users to browse resources using the MeSH hierarchy provides functionality for users as well information about patterns of use for digital librarians. Although designed for periodical articles and books, the MeSH taxonomy functions with Internet resources. This analysis of MeSH terms demonstrates user preferences that would otherwise remain obscured. More specific MeSH terms are more sensitive than more generic MeSH terms in elucidating user preferences.

Acknowledgements

The authors would like to thank the library services at St. George's, University of London for help and support with MeSH indexing.

References

[1] Robinson J, de Lusignan S, Kostkova P. The Primary Care Electronic Library (PCEL) five years on: open source evaluation of usage. *Accepted for publication* Informatics in Primary Care September 2005.
[2] WHO | International Classification of Diseases (ICD) 2006, viewed 4 January 2006, <http://www.who.int/classifications/icd/en/>
[3] SNOMED International 2006, viewed 4 January 2006, <http://www.snomed.org/>
[4] Robert B. Allen. Two digital library interfaces that exploit hierarchical structure. In *Proceedings of DAGS95: Electronic Publishing and the Information Superhighway*, Boston, MA, May 1995.
[5] Jamoulle M. Euro-Med Data Survey. URL: http://www.ulb.ac.be/esp/emd/classifications.htm
[6] Mann T. Library Research Model:A guide to classifications, cataloging and computers. Oxford; OUP, 1993.
[7] *International Statistical Classification of Diseases and Related Health Problems 10th Revision*, 2003, World Health Organisation, viewed 4 January 2006, <http://www3.who.int/icd/vol1htm2003/fr-icd.htm>.
[8] Janssen F, Kunst AE. ICD coding changes and discontinuities in trends in cause-specific mortality in six European countries, 1950-99. Bull World Health Organ. 2004;82(12):904-13.
[9] *About the UMLS Resources*, 2004, National Library of Medicine, viewed 4 January 2006, <http://www.nlm.nih.gov/research/umls/about_umls.html>.
[10] Giannangelo, Kathy, and Berkowitz, Lyle, 'SNOMED CT Helps Drive EHR Success.', 2005, *JAHIMA* 76;4:66-67, viewed 4 January 2006, <http://library.ahima.org/xpedio/groups/public/documents/ahima/pub_bok1_026463.html>.
[11] Robinson J, de Lusignan S, Kostokova P, Southgate L. Developing a digital library: The evolution of the Primary Care electronic Library (PCEL). Acta Informatica Medica 2005 13;4:31-37.
[12] MeSH Tree Structures–2005,NLM,viewed 15 May 2006.<http://www.nlm.nih.gov/cgi/mesh/2005/MB_cgi>
[13] *The |STAT Handbook*, 2002, viewed 4 January 2006, <http://www.cog.brown.edu/~max/cg144/class_files/pipehandbook.htm>.
[14] *Medical Subject Headings - Files Available to Download*, 2005, National Library of Medicine, viewed 4 January 2006, <http://www.nlm.nih.gov/mesh/filelist.html>.
[15] Specific classification of resources reveals more about user preferences in a digital library, 2005, viewed 4 January 2006, <http://www.pcel.info/mie2006/>.
[16] Athey, D 2004, 'Data Classification Using a Digital Taxonomy', *ColdFusion Developer's Journal*, viewed 4 January 2006, <http://www.paperthin.com/news/upload/CFDJ-November2004-Article-taxonomy.pdf>.
[17] Adams KC 2000, 'Immersed in Structure: The Meaning and Function of Taxonomies', *Internetworking*, vol. 3, no. 2, viewed 4 January 2006, <http://www.internettg.org/newsletter/aug00/article_structure.html>.
[18] National Library of Medicine, MeSH Browser, 2006, viewed 4 January 2006 http://www.nlm.nih.gov/mesh/MBrowser.html
[19] *Organising Medical Networked Information (OMNI)*, 2006, viewed 4 January 2006, <http://omni.ac.uk/>.
[20] *The Resource Discovery Network: evaluation report*, 2002, viewed 4 January 2006, <http://www.rdn.ac.uk/publications/evaluation/evalreport02.pdf>.