The key role of classification and indexing in view-based searching

A Steven Pollitt
Centre for Database Access Research
University of Huddersfield
UK

ABSTRACT

The application of classification schemes and thesauri to improve online information retrieval can be traced back to the beginnings of online searching itself, but the true potential for using knowledge structures in the user interface has yet to be realized. View-based searching seeks to exploit the classified arrangements in thesauri and existing classification schemes to improve the performance of such systems. HIBROWSE for EMBASE is a system which demonstrates the power of applying an approach to information retrieval which is strongly related to faceted classification. It does this by employing a point and click user interface with mutually constraining views utilising knowledge structure hierarchies for both query specification and the presentation of results. The relevance of this approach to library OPACs is discussed in the context of the digital library, concluding that our legacy of research in classification and indexing is more relevant than ever in the design of systems to cope with the problems of information access.

1. Introduction

The use of classification schemes in computer interfaces to information retrieval systems dates back at least to 1968 to a system called AUDACIOUS, developed at Syracuse University. This system used the Universal Decimal Classification to search a file of nuclear science literature (Cochrane 1982). Pauline Cochrane's suggestion in 1982 that it would seem "a horrible waste" not to use all the intellectual effort invested in classification in online retrieval systems, having demonstrated systems for doing so using UDC and DDC, seems not to have been heeded by the designers of library OPACs.

> "More than a decade has gone by since the first results came in about user difficulties with online public access catalogs (OPACs) … had we taken these findings seriously and begun the suggested improvements then, maybe we would not find the situation still as bad as ever. In our opinion, the technologies at our disposal at the time were not sufficiently advanced to provide the functionalities required. What is frustrating now is to find that system designers today, with better technologies, have learned little or nothing from those early OPAC user studies, from the analytical papers on preparing classification and thesauri for use online … from the early attempts to mount thesauri and classification systems into retrieval/search systems .." (Atherton Cochrane & Johnson 1996)

Charles Hildreth (1991) reported the results of an experiment which demonstrated an increase in recall (from 12.8% to 23.3%) for OPAC users who used a "BOOKSHELF BROWSING" feature over those relying on keywords. The increased recall was accompanied by an increase in quality of the citations retrieved, quality scores increasing from 10.36 to 18.38 as judged by experts, and a small increase in precision.

The advantages of searching from controlled vocabulary indexing using a thesaurus are clearly described by Jean Aitchison and Alan Gilchrist (1987) and Dagobert Soergel (1994), even so the ANSI and ISO Search and Retrieval protocols "contain as yet no explicit thesaurus-aided search capability, largely because of the difficulty in agreeing on the way in which thesauri can be used to enhance the search process" (Davies 1996).

The work on view-based searching described below has demonstrated how we can use a faceted thesaurus to both specify the subject matter of a query and present the results of that query through implicit searching of the underlying database. In so doing view-based searching realizes a latent potential for faceted classification, as recommended by Ranganathan in the 1920's (Ranganathan 1965), and repeated by the Classification Research Group (CRG) in the 1950s (Classification Research Group 1957).

A view-based system is a tool intended to meet the needs identified by the CRG

> "An index, a classification, an automatic selector, or any other system of 'information retrieval', is a working tool designed to help the user to find his way about the mass of published information relating to a certain field of knowledge. The user may have a detailed understanding of the pattern of knowledge in the subject he explores, or he may have only an uncertain and confused understanding of it. An information retrieval system should be designed, first, to help even the ignorant user to pass from the vague formulation of a subject in his mind to its precise formulation in the system; and then, having reached this precise formulation, to direct the searcher forward to literature references relating to it." (Classification Research Group 1957).

with the benefits brought by advancing technology of rapid processing to provide a highly interactive discovery system.

This paper takes up the cause of faceted classification and indexing as it relates to the provision of views for the OPAC searcher.

2. The Background to View-based Searching

Research and development into information retrieval in the Centre for Database Access Research at the University of Huddersfield has focused on incorporating thesauri in the retrieval system interface. The earliest system, CANSEARCH, used extracts from MEDLINE's Medical Subject Headings (MeSH) made available to cancer therapy clinicians via a touch screen terminal, taking advantage of our superior capacity for recognition over specification at the user interface. The system had to be easy to use without sacrificing on retrieval performance. CANSEARCH applied an expert systems approach, and used a rule base to both control the interaction and generate legitimate search statements on behalf of the clinician (Pollitt 1987).

CANSEARCH was an attempt at disintermediation and produced promising results. Difficulties experienced in transferring these principles into the subject area of biotechnology and biochemical genetics forced a change in approach. The expensive rule-base was removed but the structured controlled vocabulary at the interface remained. This new approach, MenUSE (Menu-based User Search Engine), accepted selections from the vocabulary and then presented the results from several search statements combining these selections. MenUSE was applied to the complete MEDLINE database (Pollitt 1988) and subsequently to multilingual interfaces for searching INSPEC in Japanese (Li, Pollitt & Smith 1992) and the European Parliament's EPOQUE system (Pollitt et al 1993) using all official European Union languages.

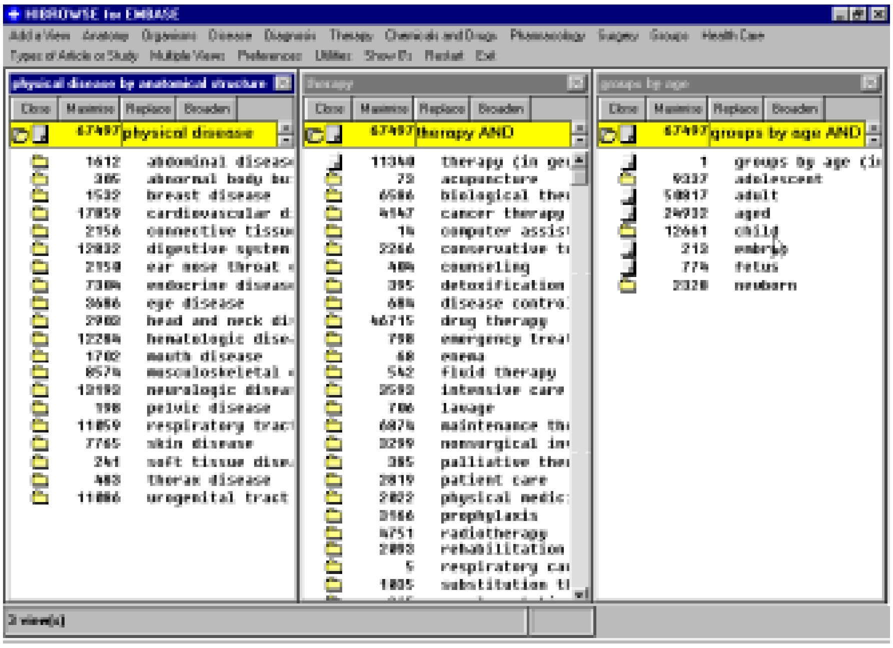
The realization that the thesaurus could be used to present results as well as a means for specifying the subject matter of a search lead to the current approach of view-based searching and the development of HIBROWSE (High resolution Interface for

BROWsing and SEarching) systems for bibliographic databases (Pollitt et al 1996).

3. HIBROWSE for EMBASE

A two year project, principally funded by the British Library Research and Innovation Centre, examined the application of usability techniques in the development of HIBROWSE for EMBASE, which accessed some 600,000 EMBASE records (Treglown et al 1997). The EMBASE database, published by Elsevier Science, comprises some 7 million records which reference the biomedical literature. The EMTREE thesaurus has 38,000 preferred terms or phrases and a vocabulary of 300,000 entry terms. The thesaurus is divided into 15 facet hierarchies. The following screens present the major features of the view-based approach in the context of retrieval from a large bibliographic database (this example was first presented in Pollitt 1997).

In Figure 1. the user has requested a presentation of three views from the EMTREE thesaurus concerning Diseases, Therapy and Groups by Age. This presentation relates to a faceted classification which takes the form DISEASES:THERAPY:GROUPS BY AGE. The system has searched for records which contain at least one of the terms in each view, 67,497 out of the 600,000 in the EMBASE subset e.g. there are 17,859 records which contain at least one cardiovascular disease term, one therapy term and one group by age term.



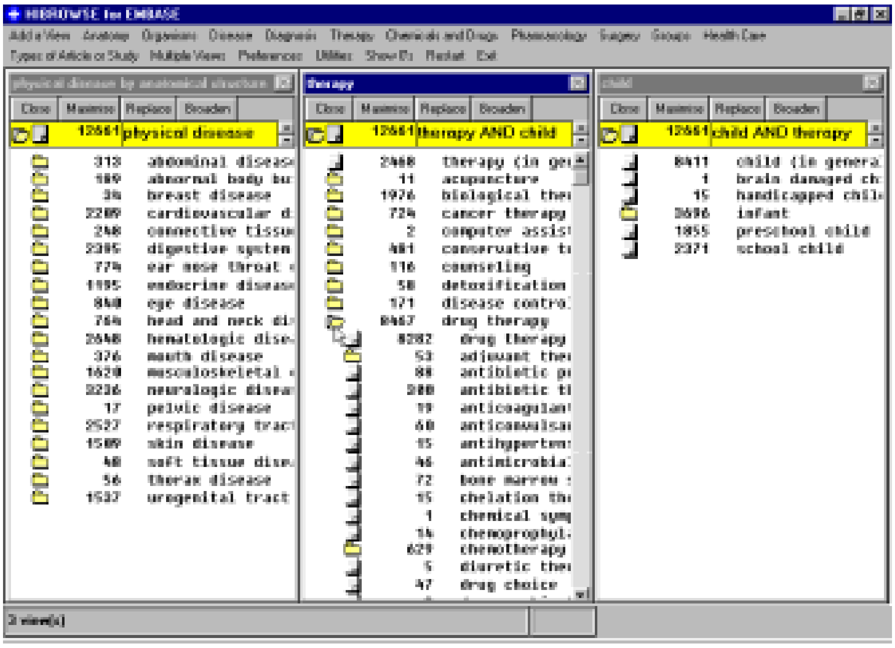
67497	physical disease
1612	abdominal disease
305	abnormal body bu
1532	breast disease
17859	cardiovascular d
2156	connective tissu
12832	digestive system
2150	ear nose throat
7304	endocrine disease
3686	eye disease
2903	head and neck di
12284	hematologic dise
1702	mouth disease
8574	musculoskeletal
13192	neurologic disea
198	pelvic disease
11059	respiratory tract
7765	skin disease
241	soft tissue dise
403	thorax disease
11086	urogenital tract

67497	therapy AND
11340	therapy (in ge
73	acupuncture
6586	biological the
4147	cancer therapy
14	computer assis
2266	conservative t
404	counseling
395	detoxification
604	disease contro
46715	drug therapy
798	emergency trea
60	enema
542	fluid therapy
3503	intensive care
706	lavage
6074	maintenance th
3299	nonsurgical in
385	palliative the
2819	patient care
2022	physical medic
3166	prophylaxis
4751	radiotherapy
2093	rehabilitation
5	respiratory ca
1805	substitution t

67497	groups by age AND
1	groups by age (i
9337	adolescent
50017	adult
24932	aged
12661	child
213	embryo
774	fetus
2320	newborn

Figure 1. Views of Disease by Site, Therapy and Groups by Age

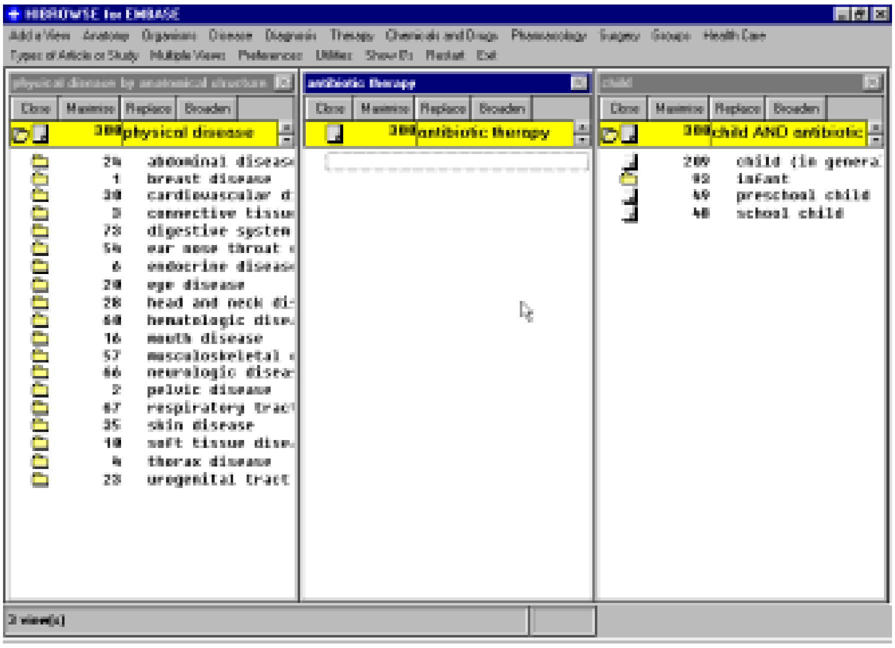
Figure 2 shows the result of two actions by the user. In the first the user refines the views by pointing and clicking with a mouse device on the term **child**. The 12,661 records are then presented according to the disease and therapy views. In the second the therapy view is expanded when the user selects the folder against **drug therapy**.



```
12661 physical disease        12661 therapy AND child          12661 child AND therapy
  313 abdominal disease        2468 therapy (in ge              8411 child (in genera
  189 abnormal body bu           11 acupuncture                    1 brain damaged ch
   34 breast disease           1976 biological the                15 handicapped chil
 2209 cardiovascular d          724 cancer therapy              3696 infant
  248 connective tissu            2 computer assis              1855 preschool child
 2385 digestive system          481 conservative t              2371 school child
  774 ear nose throat           116 counseling
 1195 endocrine disease          58 detoxification
  848 eye disease               171 disease contro
  764 head and neck di         8467 drug therapy
 2648 hematologic dise              8282 drug therapy
  376 mouth disease                   53 adjuvant the
 1628 musculoskeletal                 88 antibiotic p
 2236 neurologic disea               388 antibiotic t
   17 pelvic disease                  19 anticoagulan
 2527 respiratory tract               68 anticonvulsa
 1509 skin disease                    15 antihyperten
   48 soft tissue dise                46 antimicrobia
   56 thorax disease                  72 bone marrow
 1537 urogenital tract                15 chelation th
                                       1 chemical sym
                                      14 chemoprophyl
                                     629 chemotherapy
                                       5 diuretic the
                                      47 drug choice
```

Figure 2. Views of Disease by Site, Therapy and Child

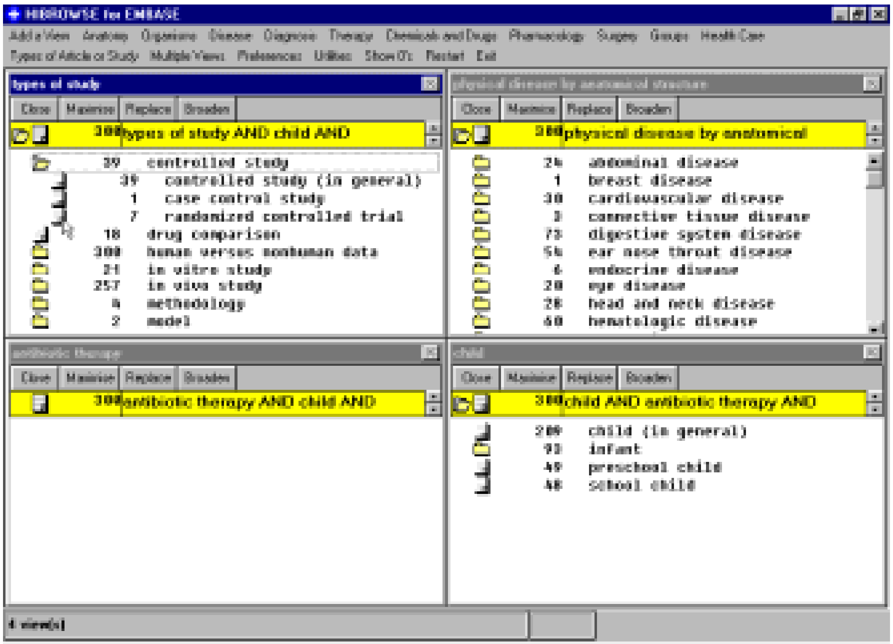
Figure 3. is the result of the user selecting **antibiotic therapy** from the expanded therapy view. All three views now concern antibiotic therapy, disease and child.



```
388 physical disease          388 antibiotic therapy          388 child AND antibiotic
 24 abdominal disease                                         209 child (in genera
  1 breast disease                                             92 infant
 38 cardiovascular d                                           49 preschool child
  3 connective tissu                                           48 school child
 73 digestive system
 54 ear nose throat
  6 endocrine disease
 28 eye disease
 28 head and neck di
 68 hematologic dise
 16 mouth disease
 57 musculoskeletal
 66 neurologic disea
  2 pelvic disease
 67 respiratory tract
 35 skin disease
 18 soft tissue dise
  4 thorax disease
 23 urogenital tract
```

Figure 3. Views of Disease by Site, Antibiotic Therapy and Child

Figure 4. shows the introduction of a fourth view on **types of study** and an expansion of **controlled study**. The user is about to view the 7 records on randomized controlled study, physical disease, antibiotic therapy and child by selecting the document icon.



```
300 types of study AND child AND
 39   controlled study
   39   controlled study (in general)
    1   case control study
    7   randomized controlled trial
 18   drug comparison
300   human versus nonhuman data
 21   in vitro study
257   in vivo study
  4   methodology
  2   model
```

```
300 physical disease by anatomical
 24   abdominal disease
  1   breast disease
 30   cardiovascular disease
  3   connective tissue disease
 73   digestive system disease
 56   ear nose throat disease
  6   endocrine disease
 20   eye disease
 28   head and neck disease
 60   hematologic disease
```

```
300 antibiotic therapy AND child AND
```

```
300 child AND antibiotic therapy AND
209   child (in general)
 93   infant
 49   preschool child
 48   school child
```

Figure 4. Views of Types of Study, Disease by Site, Antibiotic Therapy and Child

The records are retrieved by post-coordination with no guarantee that the randomized controlled trial concerned antibiotic therapy or child or a particular disease. To improve on the accuracy of the retrieval requires pre-coordination on the part of the database producer.

4. HIBROWSE for OPACS

Classification schemes for libraries have endeavored to provide a helpful arrangement for a community of users through the adoption of notations which provide a relative location for each book.

> "... it is the duty of documentalists to spread the multi-dimensional universe of knowledge along one line. We must make a linear spectrum of it. .. we have to map an n-dimensional space on a one-dimensional space. This is the problem in the organization of information and knowledge for use." (Ranganathan 1965 p. 198)

The resulting arrangement will, by compromising to a single sequence, scatter items relevant for a given need into separate locations in the library. This is the case regardless of the scheme employed. In a faceted classification the sequence of facets determines the physical arrangement with items identified by values in the final facet most widely distributed throughout the entire library.

A significant amount of time, effort and resource is devoted to the shelving of books into the sequence defined by classification code. At the University of Huddersfield there are 22 shelving assistants who spend a total of 230 hours each week between September and June replacing some 525,000 books from a stock of 338,000 unique titles, close to a million titles. Each assistant walks some 55 miles each week. Multiplying these figures by 150 will provide some measure of the resources being deployed in UK Universities. As we progress to integrate information systems and hold more diverse multimedia materials in the library OPAC, especially in schools, colleges and universities where more of these materials will be held in electronic form, the benefits from browsing the bookshelves will diminish. Resources devoted to maintaining the order can be redirected to adding value to the learning resource system as a whole.

The OPAC will provide a superior means for browsing than could be provided in a physical arrangement of items on shelves. The arrangement can be dynamic and selective to suit the changing needs of the user. The role for classification remains, but it's application can now be focused on issues of presentation at the user interface and mechanisms for searching the databases beneath it. Each of the views in HIBROWSE represents an arrangement. At the same time the selection of new views and the navigation within views can provide an automatic filtering device to exclude those items which don't possess an attribute required of each item by the user. The user can relax or impose constraints on retrieval by simple selection.

In this context faceted schemes can be extended to address the attributes of material that help to determine relevance. For example, whether a book is more suitable for a first year undergraduate than a researcher with a doctorate is not discernible from most OPACs, yet this is a crucial criteria understood by all users and can be readily applied to avoid wasting time examining the item. The length of a video or the time anticipated for a Computer Based Learning program to run is an important characteristic or facet when the user is faced with a choice.

5. Conclusion

The expectations of future library users will be high given the power and sophistication of computer games. Responding to these expectations will need to relate the technology of today with the thinking demonstrated by those working in the difficult field of classification over many years. The phenomena of the World Wide Web with point and click, global navigation and incredibly powerful search engines has not provided a solution as yet but often leads to frustration for middle and high school students as identified by Elliot Solloway and Raven Wallace (Solloway & Wallace 1997).

> "The ease with which students and teachers learn to use Web browsers is truly astounding… No manual was needed; just point and click and the technology disappeared and activity flowed forth. But using a browser is not the same as "productively navigating the Web". Even the non-linear, videogame generation becomes quickly frustrated and stymied. … The Web is not a Library.. A library contains comprehensive collections of information resources. While the Web contains collections, there is no guarantee of comprehensiveness, of systematically covering the human record. A library is a purposely and purposefully constructed organization: but, by design, the same cannot be said of the Web."

The advent of new technology tempts some systems developers into thinking that techniques devised in the days before the technology are no longer relevant. "There are still those who would have us believe that natural language queries and ranked output offer the panacea for end-user searching. The size of result sets, and the potential for organizing these according to different views, suggests a continuing role for classification as de Grolier stated some thirty years ago" (Pollitt et al 1996).

> "We feared some years ago that classification was becoming useless, that the treatment of natural language texts by machine … would replace classification. Classification and the classificationists would become something like the dinosaurs, killed by the progress of evolution. This has proved to be a complete fallacy. When you examine the new literature you find that more and more classification … is seen as something quite essential in information retrieval … it is quite evident that hierarchies, generally speaking, are something which cannot be avoided in an information retrieval system which is to be

useful for the reader." (de Grolier 1965)

The role of classification and indexing is crucial to our continuing development and systems which take advantage of the evolving ways to represent knowledge structures at the user interface will ensure a higher quality and more rewarding experience in our life-long learning.

Acknowledgements

Thanks are due to Patrick Braekevelt, Martin P Smith and Geoffrey P Ellis, the team that developed HIBROWSE, and to Mark Treglown and Janet E Finlay, the researchers who provided input on usability to the design process. We are grateful to the British Library Research and Innovation Centre which provided the major part of funding for the HIBROWSE for Bibliographic Databases project (grant RDD/G/282) together with Elsevier Science BV, Software AG (of the UK), Sun Microsystems, the University of Huddersfield and Elonex, who provided additional sponsorship, and the Medical and Dental Library at the University of Leeds for collaboration.

References

Aitchison, Jean and Gilchrist, Alan (1987) *Thesaurus Construction. 2nd Edition.* Aslib London 1987.

Atherton Cochrane, Pauline and Johnson, Eric H. (1996) Visual Dewey: DDC in a Hypertextual Browser for the Library User *in* Rebecca Green (ed*) Knowledge Organization and Change, proceedings of the Fourth International ISKO Conference 15-18 July 1996 Washington DC, Advances in Knowledge Organization vol. 5* (1996) INDEKS VERLAG pp 95-106

Cochrane, Pauline A. Classification as a User's Tool in Online Public Access Catalogs *in* Dahlberg, Ingetraut (ed) *Universal Classification I: Subject Analysis and Ordering Systems Proceedings of the 4th International Study Conference on Classification Research Augsburg 28 June - 2 July 1982* Indeks Verlag pp260-268.

Classification Research Group (1955) The need for a Faceted Classification as the basis of all methods of information retrieval. A memorandum first published as UNESCO (IAC Doc. Ter. PAS) document 320/5515, 26 May 1955 reprinted in *Proceedings of the International Study Conference on Classification for Information Retrieval,* Dorking, May 1957 London: Aslib, pp. 137-147

Davies, Ron (1996) Thesaurus-aided Searching in Search and Retrieval Protocols *in* Rebecca Green (ed*) Knowledge Organization and Change, proceedings of the Fourth International ISKO Conference 15-18 July 1996 Washington DC, Advances in Knowledge Organization vol. 5* (1996) INDEKS VERLAG pp 137-143.

de Grolier, Eric (1965) Current trends in theory and practice of classification. *in* Pauline Atherton (ed) *Classification Research - Proceedings of the Second International Study Conference,* Elsinore Denmark 14-18 September 1964, FID/CR Copenhagen pp 9-14

Hildreth, Charles R. (1991) End Users and Structured Searching of Online Catalogues: Research Findings, , *in*: Fugmann, Robert, *ed. Advances in Knowledge Organization: Tools for Knowledge Organization and the Human Interface Vol 2. Proceedings of the 1st International ISKO Conference, Darmstadt, 14-17 August 1990.* Frankfurt/Main: INDEKS VERLAG, pp. 9-24.

Li, Chung Sheng, Pollitt, A. Steven, Smith, Martin P. (1992) Multilingual MenUSE - A Japanese front-end for searching English language databases and vice versa *in* Tony McEnery & Chris Paice (eds). *14th BCS IRSG Information Retrieval Colloquium,* Lancaster Pub: Springer Verlag pp 14-37

Murphy, Frances J., Pollitt, A. Steven, White, Philip R. (1991) *Matching OPAC user interfaces to user needs - Final Report of the British Library Project,* BLR&DD

Report No: 6041 The Polytechnic of Huddersfield, April 1991

Pollitt, A. Steven, (1987) CANSEARCH: An expert systems approach to document retrieval *Information Processing and Management*, Vol. 23, no 2, pp 119-138, 1987

Pollitt, A. Steven, (1988) A common query interface using MenUSE - A Menu-based User Search Engine *12th International Online Information Meeting*, Vol. 2, pp 445-457, London, Dec. 1988

Pollitt, A. Steven, Ellis, Geoffrey P., Smith, Martin P. Gregory, Mark R, Li, Chun Sheng and Zangenberg, Henrik (1993) A common query interface for multilingual document retrieval from databases of the European Community Institutions *Proceedings of the 17th International Online Information Meeting,* London, England. Learned Information December 1993 pp 47-61.

Pollitt, A. Steven, Smith, Martin P., Treglown, Mark and Braekevelt, Patrick (1996) View-based searching systems - progress towards effective disintermediation *Proceedings of the 20th International Online Information Meeting*, London, December 1996 Learned Information Limited, pp 433-446

Pollitt, A. Steven (1997) Interactive Information Retrieval based on Faceted Classification using Views in Knowledge Organization for Information Retrieval. *Proceedings of the 6th International Study Conference on Classification Research*, University College London 16-19 June 1997 (to appear)

Ranganathan, S.R. (1965) *A descriptive account of Colon Classification.* Bangalore: Sarada Ranganathan Endowment for Library Science

Soergel, Dagobert (1994) Indexing and Retrieval Performance: the logical evidence. *Journal of the American Society of Information Scientists*, 45(8), 1994 pp 589-599.

Soloway, Eliot and Wallace, Raven (1997) Does the Internet Support Student Inquiry? Don’t ask. *Communications of the ACM* May 1997 Vol 40 No. 5 pp 11-16

Treglown, Mark, Pollitt, A. Steven, Smith, Martin P., Braekevelt, Patrick, Finlay, Janet E. (1977) *HIBROWSE for Bibliographic Databases: a study of the application of usability techniques*. British Library Research and Innovation Report No. 52. British Library/University of Huddersfield (to appear)