International Trends in Subject Analysis Research

I.C. McIlwaine / N.J. Williamson

Ia McIlwaine is Professor of Library and Information Studies in the University of London and Director of the school of Library, Archive and Information Studies at University College London, where she teaches Knowledge Organization and Information Sources. She is Chair of the IFLA Section on Classification and Indexing and FID/CR and a member of the ISKO Scientific Advisory Council. She is Editor in chief of the Universal Decimal Classification.

Nancy Williamson is Professor Emeritus in the Faculty of Information Studies, University of Toronto where she teaches a course in the Subject Approach to Information. She is a Vice-President and member of the Executive Board of ISKO. Her recent areas of research include a feasibility study on converting UDC to a fully faceted system using Class 61 Medical Science, organization of the internet and the problems of interdisciplinarity and traditional classification systems.

I.C. McIlwaine / N.J. Williamson (1999). **International Trends in Subject Analysis Research** *Knowledge Organization, 26(1),* 23-29. 14 refs.

ABSTRACT: This paper describes a survey of subject analysis research over the ten year period 1988 to 1998. Data are drawn from the "research environment" encompassing publications, conference papers, major bibiographic resources in the field of Library and Information Science and selective searches of the Internet. Findings reveal major and minor areas of research activity. Trends and developments are identified and conclusions drawn. Strengths and weaknesses in the approaches taken to subject analysis research are discussed and suggestions for improvements are made with a view to future research directions.

1. Introduction

Subject analysis can be defined as the process whereby documents, data and other information carriers are described and represented according to their subject content. The recorded result of this mental process may be manual or machine-based. The documentary languages used in the representation may be systematic and denoted by symbols (i.e. classifications) or alphabetic systems of controlled vocabulary. In recent years this field has been referred to as "knowledge organization". The area of research is vast and involves theoretical, experimental and applied methods of investigation.

This paper[1] focuses on "subject analysis research" as it is represented in the published output of the field over the ten year period, 1988 to 1998. Since the volume of publication is huge, the observations and findings in this study have been limited in two ways - by time frame and by restricting the examination of research papers to the sources recognized as the most important in the field. Within these limitations, an attempt has been made to obtain a profile of the research and to identify some trends and potential future directions. In order to make the task more manageable, the authors decided to concentrate on three areas: universal classification systems, thesaurus design and development and efforts to organize the Internet. These have been examined in slightly more detail than other aspects of subject analysis.

In North America, 1988 was the year in which a conference was held at SUNY Albany on *Classification Theory in the Computer Age: Conversations Across the Disciplines* (1989). While the central rôle of classification in information retrieval was already well established, this conference raised particular concerns for the future of classification as it had hitherto been viewed. Meanwhile, on the other side of the Atlantic, in the UK, Brian Vickery, a founder member of the Classification Research Group (CRG) and pioneer in classification theory and research in information retrieval, was being honoured. The *Journal of Documentation* produced a special issue entitled "Essays presented to B.C. Vickery" (1988) to mark his retirement as Professor of Library Studies in the University of London three years previously.

The CRG is arguably one of "the most potent influences in the development of classification theory" (Foskett 1971, p. 210) and the Vickery Festschrift brought into focus the accomplishments of the past and set the tone for the future. By strange coinci-

dence, the other end of the ten year period, the year 1998, marked the 50th anniversary celebrations in London of the Conference on Scientific Information held by the Royal Society which created the framework for the UK's postwar library structure. A side effect was the setting up of a small group of experts to investigate the organization and retrieval of scientific information, which eventually became the CRG and was the impetus for its early research - research which is fundamental to subject analysis, even today. Between 1988 and 1998 a number of other significant events had taken place. These included the Ranganathan Centenary in 1992, the founding of the International Society for Knowledge Organization (ISKO) and its biennial conferences in 1989, the continuance of the series of FID/CR International Study Conferences with two further meetings held in 1991 and 1997, the Allerton Institute on "New Roles for Classification in Libraries and Information Networks" (Carter, 1995), and the nine ASIS SIG/CR Classification Research Workshops which began in 1990. It was also a period of major advances in the design of OPACs, innovations in the design of information systems generally and the application of computer technology to information retrieval. Above all, it witnessed the dawn of the age of the Internet. It was a period of highly active expansion in research and publication, of exchange of information and tremendous creativity. What does the research literature tell us about what happened and what conclusions can we draw from it regarding the prospects for the next century?

2. Sources

In this study it was not our intention to enter into a discussion of "what is research". However, in defining the procedures it was necessary to make some arbitrary decisions in order to seek out pointers for the way ahead and the paths that future research might follow. We have defined our remit as comprising those enterprises whose products can be found as the publications and presentations offered in what professes to be the "research environment" of the field. We have drawn on the major research journals, important research-oriented conferences, various library and information science bibliographical databases and selected searches of the Internet[2]. Even with an exhaustive search of these sources, there is reason to believe that there are relevant major projects carried out in private corporations and other organizations for which reports are never published and which are never written up in the literature of knowledge organization. Fortunately, a few of these do find their way to the conferences - it is highly regrettable that more do not.

3. Methodology

The scanning of these sources resulted in a total of 575 items which were selected as being relevant to the purpose of the study. In the first stage of the analysis these were grouped under 20 headings (with some falling under more than one head). These categories were then used to determine the most active areas of research. In a second step the papers were also organized by date of publication in order to gain some insights into the growth and development of research in particular subject areas.

4. Findings

The largest number of papers was located in the following categories: universal classification systems, cognitive processes, thesauri, structure and relationships, terminology and natural language processing (including cluster analysis, semantic classification, and automatic indexing). Smaller groupings of papers included works on concepts and categories, semantics, semiotics, linguistics, classification of images, taxonomy and ontologies. Most of the research was experimental in nature with less emphasis on applied and theoretical research. The diversity of topics has increased over time. Some topics were spread evenly over the ten years, others appeared in "clusters". To some extent this is explicable by two factors - the appearance of "special issues" of journals and the publication of conference proceedings.

Both FID/CR and ISKO routinely develop their conferences around central themes, with a broad general title that enables the assimilation of an equally broad range of papers, accommodating many diverse topics. However, it is significant that the second ISKO conference in 1992 was heavily weighted on the side of cognitive theory and process, in line with the theme "Cognitive Paradigms in Knowledge Organization" (Neelameghan *et al.*, 1992). At the 1991 FID/CR International Study Conference (Hudon and Williamson, 1992) one of the most frequently discussed topics was thesaurus design, and this was to become a feature of successive similar events, as the trend for word-based retrieval systems in an online environment has come virtually to dominate current thinking.

The fifth International ISKO Conference (Mustafa el Hadi *et al.*, 1998) placed particular emphasis on structure and relationships, but in that context there were three strong sub-topics - cognitive approaches, linguistic aspects and the design of information systems. Both *Knowledge Organization* and *Cataloging & Classification Quarterly* had several theme issues. There may be other reasons for these "blips" in the pattern of research, such as the locale and cultural background of a particular conference venue, but this

is not readily apparent. It is also possible that the primary concerns of researchers at a particular time have some effect, and the changes in technology over the decade under review certainly also had a rôle to play.

Examination of the published research indicates that what began as a focus on bibliographic classification in the traditional sense has opened up considerably with respect to scope, content and application. The universal classification systems appear at every conference, signifying their ongoing relevance as they attempt to adapt themselves to the ever-changing environment. However, as the background of conference participants becomes increasingly diversified, these systems, in their pure sense, occupy a diminishing percentage of conference time. It is noteworthy that a recent issue of *Library Trends* (Bowker and Star, 1998) devoted to classification included no article devoted specifically to any of the major general schemes. Nevertheless, the theory and principles of classification remain fundamental to all aspects of knowledge organization and, in this respect, their presence is stronger than ever; for example, faceted classification is being given progressively greater attention, and now permeates research in many areas of knowledge organization.

Research into thesaurus design and construction is one of the most notable examples of this. Indeed, the fundamental importance of classification and classificatory structure in all kinds of information systems seems now to be clearly understood. This would not have been true ten years ago. Terminology has always been a fundamental factor in the theory and design of information structures, but its rôle in the research of knowledge organization is growing. With the development of increasingly sophisticated methods of text analysis, semantics, semiotics and linguistics are regularly seen as integral elements of knowledge organization. Other topics, for example the Internet and undiscovered public knowledge, have come to the fore more recently - an indication of increased interdisciplinary directions in the research. Some areas of investigation are not as new as they may seem. Hypermedia and frame-based systems were there in 1988, but their prominence in discussions has increased with the growing sophistication of technology. The use of the term "knowledge organization" as a concept in the literature is tangible evidence of changes in the perspectives and output of the field and appears to coincide very closely with the change of name of the journal *International Classification* to *Knowledge Organization* in 1993.

Most of the published research is still highly motivated by academia. There are many reasons for this, not least of which is the increase in accountability for those who teach in universities to "publish or perish". The majority of the authors in the survey are from the schools and departments of library, archive and information studies, with a slow increase in participation from other disciplines. A small number of practitioner-researchers from national and other large libraries and a small, but growing, number of consultants and researchers from commercial firms have participated in the conferences of the past five years.

With few exceptions, there has been little team effort. This has some explanation in the comparatively new status of the discipline, its origin very often within Faculties devoted to the Humanities and its struggle to achieve recognition within the traditional university sphere. It now sees itself as a science (an arena where team-based research is the norm) but it still suffers from growing pains and clings to its humanistic roots. Many researchers are working alone and it is difficult to find cumulative research which builds on the work of others. As a result, much of the output from conferences focuses on small projects of individual interest.

In addition to the nature and origins of the discipline, there is frequently a lack of funding for major projects. However, the diversification of participants and the participation of the consultants is an encouraging sign of the forging of important new relationships among researchers. But the fact that there is very little money being poured into research of this kind remains a major obstacle. The exceptions are research being carried out by large corporations in conjunction with the development of their own information systems and companies in the "information business", most notably OCLC and its Forest Press Division.

5. The Universal Classification Systems

Aside from ongoing revision and updating, perhaps the greatest progress made in the universal classification systems since 1988 has been the conversion of the three most-used systems, DDC, UDC and LCC, into machine-readable form. Now, in the interest of enhancing retrieval, the focus is on making them adaptable to the computerized environment in the most useful way possible and, in particular, in adapting them for use as retrieval tools for information on the Internet.

The users and supporters of UDC have long debated its qualities as a computer-manipulable system. To this end its flexibility and its numerical notation are among its greatest assets. In its overall development UDC is becoming "more faceted" (McIlwaine, 1994, p.12) in nature. A major project (McIlwaine and Williamson, 1993) currently under way is a feasibility study for restructuring UDC into a fully faceted system. Class 61 Medical Sciences is being restructured

and updated, using the framework from Class H of the Bliss Bibliographic Classification. It is hoped that the main 61 tables will be completed in the near future. While the final impact of the study is not yet known, its existence has already had considerable influence on the ongoing revision of UDC.

With some research and considerable effort, the Library of Congress Classification is now in machine-readable form. Many were sceptical that this could happen, but it was not a question of whether LCC could be converted to machine-readable form; rather it was whether the result would be intelligible and useful. Fortunately that mission has been accomplished successfully, and in "Classification Plus" for the first time, we have the facility to link LCC and LCSH - something DDC is also attempting, but not yet comprehensively. We look forward to the completion of the project and the publication of all the classes of the scheme on the CD-ROM in what, especially linked with "Catalogers' desktop", is a most useful tool both for the practitioner and the teacher. From an international point of view, an important requirement in the LCC conversion was the development of the USMARC for classification data. The result is a format of international importance which has in turn spawned a UNIMARC format usable with all of the major systems. This is still in draft form at present, but should be ready for publication after the Bangkok IFLA meeting in August 1999.

The most intensive and cohesive research programme is the one being carried out at OCLC in conjunction with the Dewey Decimal Classification (DDC). Early research on OPACs carried out by Karen Markey Drabenstott (Markey and Demeyer, 1985) did much to confirm the importance of classification and its use as a retrieval tool in online catalogues. This opened the door to exploration and innovation in the design, maintenance and development of classification schemes and the design of online catalogues. The result has been such products as "Dewey for Windows" and the use of DDC in "NetFirst", a catalogue of Internet resources. The work of Diane Vizine-Goetz (1998) on the adaptation of DDC as an Internet subject guide should be referred to in this connection. She is developing the classification as a resource for searching the Internet through adding supplementary terminology, revising the captions in the scheme so that their currency and expressiveness are increased and ultimately developing a prototype Web-accessible Dewey Subject Guide. Another OCLC-related project that should be mentioned in passing is the research going on at the University of Huddersfield in the UK under the direction of Steven Pollitt (1998) using view-based searching based on facets derived from DDC captions.

While many may criticize the traditional schemes for their 19th century heritage, it is important to note that it is precisely those schemes that have come to the rescue of the chaos of the Internet. There are some who still think that a new general classification scheme is needed. Is it? Will it come to pass? In the ten year period investigated only three papers were found which were devoted to a major discussion of this topic. The calls for a new universal classification system are rapidly becoming fainter, suggesting that this is unlikely to happen. The nearest approach to such interest is the slow but steady publication of the revised Bliss Bibliographic Classification (BC2).

6. Thesauri

It would not be easy to pick the most often discussed research topic in the ten years, but evidence suggests that the principles, practices, design and construction of thesauri would be a major contender. A perusal of the research papers identified many on this subject alone, as well as those discussing it in conjunction with other topics. The application of facet analysis in thesauri was definitely seen as an important factor in thesaurus design. Thesauri were also discussed as navigation tools, as tools in the interactive use of information systems, and as components in expert systems, to name but a few contexts.

There is also some indication that the thesaurus is a tool in transition. It has been suggested that thesauri should introduce terminological definitions into their displays and that research is required to determine whether new kinds of thesauri are needed for the online systems. Some predict that the rôle of thesauri will change and that they will be tied more closely to retrieval than to indexing. If this should come to pass, two different types of thesaurus may be needed to satisfy the two quite different requirements that are made of it; one as an "organizer" used for systematic arrangement of indexes of any type, the other as a retrieval aid which permits the searcher to use his/her own words and connect to the term used by the database being searched, in the manner of the original thesaurus of Roget.

An important matter for consideration is the multi-lingual thesaurus. In the collections of international conference papers, one might expect to find many on this topic. Surprisingly, only three items were found that were directly devoted to it, although there is evidence elsewhere that this is an important issue in Europe and in other parts of the world where it is essential to be able to access information in multiple languages. For instance, an international project that has considerable future potential was reported at the 1998 IFLA conference. This is Project MUSE (MUltilingual Subject Entry) – the creation of a

French, English and German thesaurus – a joint project between the British Library and the National libraries of France, Switzerland and Germany. To date, a pilot study has been undertaken, using sport and theatre as the selected subject fields, and terms are being extracted from the three national subject authority files (LCSH, Rameau and the German Schlagwortnormdatei (SWD)). The intention is not to translate terms from one language to another or to establish a homogeneous thesaurus with terms in different languages, but rather to link the existing national thesauri and offer the opportunity of switching to an OPAC user, or to someone accessing a national bibliography. More projects of this kind (mapping systems to one another) are clearly an important future trend, and one that needs to be considered more seriously in relation to the general classification schemes as well as subject headings lists. The "Classification Literature" bibliography published in *Knowledge Organization* indicates that there are numerous such thesauri being created and that there are serious problems to be faced by the creators of multi-lingual tools.

This leads naturally on to the question of compatibility among thesauri concerned with the same subject and in the same language. Here again, research on mapping techniques is beginning to appear - an approach to switching languages under another name. Clearly, as controlled vocabularies go, thesauri are seen as tools which will move into the 21st century, but perhaps in a different form with different uses. Also, there is now quite a number of thesauri on the Internet. Perusal of a selection of these online thesauri, some of which also exist in printed versions, identifies important differences in the methods used for their display and in the ways they can be accessed and manipulated online. This is an area that definitely needs further research and probably also revision of standards which are now becoming distinctly elderly having seen no major changes since the early 1980s.

7. The Internet

While universal classification systems and thesauri are well established topics for research, actions on the organization of the Internet are, for obvious reasons, relatively new. In this study the sources used located only five papers in which the primary focus was the problems of the organization of the Internet. All were published in the last two years covered by the study (i.e. 1997 and 1998). Although the *Journal of Internet Research* has been in existence for several years, at the time when this research was conducted it had yet to address the problems of organizing and accessing the data. So far, library and information scientists have had minimal impact on the Internet system as a whole. Nevertheless the Internet itself is the best source on information about what is happening. The three best known classification schemes occur in a number of different enterprises for the organization of websites. Analysis of some of these websites in an earlier paper (Williamson, 1997) identified the fact that the application of classification at the sites was often superficial and poorly executed, and was very varied in nature. One of the most informative websites in this area is CyberStacks(sm) (http://www.public.iastate.edu/~CYBERSTACKS/) which lists classification systems and other controlled vocabularies on the Internet. As evidence of activity this list has grown considerably over the past two years. At this site there is also a list of "Net projects" which are in progress. Among the approaches to the problems being addressed are visual browsing, citation indexing, data mining technologies, hypertext thesauri and intelligent software agents.

8. Conclusion

When the findings of this survey are analyzed, it is clear that there is considerable research activity and that there is international involvement. It is also apparent that many long standing problems still exist and new ones are appearing on the horizon. What has been achieved? Where should we go from here? In the light of the findings it is interesting to reflect on Brian Vickery's remarks at the closing of the 6th International Study Conference on Classification Research in London in June 1997. Looking back to the Dorking Conference 40 years earlier, he said "Many of the themes at that [Dorking 1957] conference have been renewed during the last few days. Here we have had papers discussing the perceived importance of classification; the effect of social conditions on the way knowledge is structured; the relevance of facet analysis; facet sequence in the build-up of compound subjects; problems of polyhierarchy: paradigmatic and syntagmatic relationships - all these issues were raised at Dorking. *Plus ça change...*" (Vickery, 1997, p. 180).

Does this have a familiar ring? It seems that this ten year survey is saying much the same thing. We have continued to debate the same problems. The interest and willingness to come to grips with those problems is there but the cumulative value of the research has eluded us. How do we meet the challenge of this shortcoming? To some extent, we have embraced interdisciplinarity and broadened our horizons over the period just past, but Vickery feels that the implications of knowledge organization have not been "fully accepted within information science" (Vickery, 1997, p.181). His recommended remedy is to look beyond our own domain and to examine more closely the "variety of ways in which public knowledge can be organized and consider the implica-

tions for information systems" (Vickery, 1997, p.181). There should be more research which focuses on the knowledge structures in the literature of different subjects. We should seek knowledge which might form a basis for generalization and a general theory of knowledge organization. This may be very wise advice indeed. The pattern of research over the past ten years strongly indicates that we may have focused too much on individual problems, without seeing them as symptoms of more universal concerns. We need to look at history. Much of the pioneering work in the 1950s and 1960s began as the development of faceted classification schemes for small specialized areas of knowledge. In the mid 1960s the CRG began work on a "New general scheme of classification", work that has never been completed and now never will be. We should treat that as an object lesson, and try to ensure that the findings of the nineties do not once again flounder in the next millennium.

Notes

[1] This paper is an edited and updated version of a presentation made by the authors at the 1998 Annual Meeting of the American Society for Information Science (ASIS) held in Pittsburgh, USA, October 28, 1998.

[2] Sources of data included research papers from *Knowledge Organization* (formerly *International Classification*), *Journal of Documentation*, *Cataloging & Classification Quarterly* (Special theme issues); activities and publications of the IFLA Section on Classification and Indexing; proceedings of the ISKO conferences, the FID/CR International Study Conferences on Classification Research and the ASIS SIG/CR workshops; the *Annual Review of OCLC Research, Extensions and Corrections to the UDC, 1993-1997;* bibliographic resources including *Library and Information Science Abstracts (LISA),* and *Information Science Abstracts (ISA).* The "Classification Literature" section of *Knowledge Organization* was selectively scanned for publications and as a source of themes evident in the literature; selected relevant Internet sites were also examined.

References and Selected Bibliography:

Bowker, G.C. and Star, S.L. (Eds.) (1998). How classifications work: problems and challenges in an electronic age. In: *Library Trends*, 47(2): 185-337.

Carter, R. (Ed.) (1995). New roles for classification in libraries and information networks: reports from the Twenty-sixth Allerton Institute, October 23-25, 1994. In: *Cataloging & Classification Quarterly*, 21(2): 3-118.

Classification Theory in the Computer Age: Conversations Across the Disciplines. (1989). Proceedings from the Conference, November 18-19, 1988, Albany, New York. Albany: School of Information and Science Policy and the Professional Development Program, University at Albany, State University of New York. Essays presented to B.C. Vickery (1988). In: *Journal of Documentation,* 44 (3): 199-258.

Foskett, A.C. (1971). *The Subject Approach to Information.* 3rd ed. London: Bingley.

Hudon, M. and Williamson, N.J. (Eds.) (1992). *Classification Research for Knowledge Representation and Organization: Proceedings of the 5th International Study Conference on Classification Research, Toronto, Canada, June 24-28, 1991.* Amsterdam: Elsevier.

McIlwaine, I.C. (1994). Report of the UDC Editor in Chief. *Extensions and Corrections to the UDC.* No. 16. The Hague, Netherlands: UDC Consortium. 9-18

McIlwaine, I.C. and Williamson, N.J. (1993). Future revision of UDC: progress report on a feasibility study for restructuring. In: *Extensions and Corrections to the UDC.* No.15. The Hague: UDC Consortium. 11-17.

Mustafa el Hadi, W; Maniez, J.; Pollitt, S.A. (Eds.) (1998). *Structures and Relations in Knowledge Organization: Proceedings of the 5th International ISKO Conference 25-29 August 1998, Lille, France.* Würzburg: ERGON Verlag.

Markey, K.; Demeyer, A.N. (1985). *Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated into the Subject Capabilities of an Online Catalog.* Dublin, OH: OCLC Computer Library Center, Office of Research.

Neelemaghan, A.; Gopinath, M.A.; Raghavan, K.S.; and Sankaralingam, P. (Eds.) (1992). *Cognitive Paradigms in Knowledge Organization: Proceedings of the 2nd International ISKO Conference, Madras, India August 26-28, 1992.* Bangalore: Sarada Ranganathan Endowment for Library Science.

Pollitt, A. S. (1998). The application of the Dewey Decimal Classification in a view-based searching OPAC. In: W. Mustafa el Hadi, J. Maniez and S.A. Pollitt (Eds.). *Structures and Relations: Proceedings of the 5th International ISKO Conference, 25-29 August, 1998, Lille France.* Würzburg: ERGON Verlag. 176-183.

Vickery, B.C.(1997). Issues in knowledge organization. In: *Knowledge Organization for Information Retrieval: Proceedings of the 6th International Study Conference on Classification Research, University College London, 16-18 June, 1997.* The Hague, Netherlands: UDC Consortium. 180-182.

Vizine-Goetz, D. (1998). Dewey as an internet subject guide. In: W. Mustafa el Hadi, J. Maniez and S.A. Pollitt (Eds.) *Structures and Relations in Knowledge*

Organization: Proceedings of the 5th International ISKO Conference, 25-29 August, 1998, Lille, France. Würzburg: ERGON Verlag. 191-197.

Williamson, N.J. (1997). Knowledge structures and the internet. In: *Knowledge Organization for Information Retrieval: Proceedings of the 6th International Study Conference on Classification Research, University College London, 16-18 June 1997.* The Hague, Netherlands: International Federation for Information and Documentation. 23-27.

Professor Ia C. McIlwaine, Director, School of Library, Archive and Information Studies, University College London, Gower Street, London WC1E 6BT, United Kingdom;
Tel: +44 0171 380 7205; Fax: +44 0171 383 0557;
E-mail: i.mcilwaine@ucl.ac.uk

Professor Nancy J. Williamson, Faculty of Information Studies, University of Toronto, 140 St. George Street, Toronto, Canada M5S 3G6;
Tel: +1 416 978 7079; Fax: +1 416 971 1399;
E-mail: williams@fis.utoronto.ca