Guest Editorial
Meeting the challenge

Brian Vickery
Oxford

1. Introduction

It is a great privilege to be asked to contribute some comments to this special issue of JIS. I have enjoyed reading these papers, and seeing how much our field has moved on since the Journal of Documentation attempted a similar historical survey in 1994 [1], just as the internet was ‘taking off’.

My first thoughts were of Jason Farradane, who played such a big part in founding the Institute of Information Scientists in 1958. The IIS has now disappeared within CILIP, and Jason would have been very dismayed at this development.

I first encountered him at the Royal Society Scientific Information Conference of 1948, where he spoke on ‘the scientific approach to documentation’. In 1952 he was one of the founder members of the Classification Research Group, where he was often defending a minority position, connected to his own ideas of relations between topics, to which Stella Dextre Clarke refers in her paper. He was a serious man, but the more I talked with him, and disagreed with him over many things, the more I liked him.

In 1955, we were invited together to a meeting of the Aslib Aeronautical Group, at which Cyril Cleverdon spoke about his plans for retrieval evaluation (Cranfield1), which Stephen Robertson discusses in his paper. We spoke about faceted classification, and Cyril challenged us to prepare a scheme for aeronautics, that he could use in his test. With inordinate bravado, we accepted the challenge, and – after much advice from Cyril and his team – managed to produce one. In retrospect – given that it was compiled by two chemists – we were amazed that it performed at all.

What else was happening around about 1958? There were printed publications of every kind. There were libraries – public, national, government, academic, institutional, industrial – and they were classified and had catalogues. We were much concerned with interlibrary loan, and Donald Urquhart had recently been told to drop all other work and start to plan a new national science lending library for the UK. There were printed indexing and abstracting periodicals, and bibliographies. An industrial firm might have, as well as a technical library, a centre for internal reports, a section dealing with patents, an ‘intelligence’ unit for commercial information, and a correspondence registry. Some of us were playing around with punched or edge-notched cards for ‘mechanized retrieval’. There were reported to be 1000 electronic computers in the USA, and 160 in Europe. Some firms were building structured computer files searched by ‘database management’ software.
An International Conference on Scientific Information was held in Washington in 1958 (ICSI [2]). Coming as it did near the beginning of the impact of the computer on information provision, it to some degree charted the future. In the titles of its papers were many of the concepts we still have today: users’ requirements, evaluation, effectiveness, efficiency, cost analysis, quantitative survey, international cooperation, codes, computers, mechanical translation, linguistics, machine searching, semantics, information retrieval. An inventive IBM engineer, Peter Luhn, distributed a computer-based ‘keywords-in-context’ index to the titles of the pre-published conference papers. For some participants this was their first encounter with a computer-based information product.

2. What is information activity?

I like to think of an analogy. There is a social activity known as medical practice, aiming to help people to become more healthy. Medical science seeks principles and methods that will improve medical practice. The practice, broadly speaking, consists of two activities: diagnosis and treatment. Diagnosis is identifying what is the most probable cause of a particular state of ill health. Treatment is deciding what action is most likely to counteract the cause. Medical science seeks to understand what are the potential causes of ill health; to develop methods of correct diagnosis; to understand and expand the range of possible treatments; and to develop methods of deciding what treatment is most likely to be effective in a particular case.

Information practice, again broadly speaking, consists of two activities that we may call diagnosis and provision. Diagnosis is identifying what is the most probable information need of a user in a particular state of information want. Provision is deciding what action is most likely to meet that need. Information science seeks to understand the potential range of user situations giving rise to information want and need; to develop methods of identifying the actual information needed; to understand and expand the range of possible ways of satisfying information need; and to develop methods of deciding what way is most likely to be effective in a particular case.

So information practice is concerned with facilitating the interaction between knowledge seekers through channels with knowledge (personal and recorded). Stepping back from practice, we may see the role of the science as exploring the characteristics of people and their ‘information behaviour’, the features of knowledge records of every kind, the variety of channels (oral, written, printed, graphic, digital) that may be used to transmit information, and how the three elements interact. Science and technology are now so closely linked that analysis and experiment lead quickly on to invention, to the introduction of new channels (and documents).

In the beginning, the role of ‘information science’ was to study the activities of the information practitioner – what he or she did, and how it might be done better. But the science indicated above goes far beyond the practitioner, to explore all the elements of the process of ‘becoming informed’. These elements are of interest to investigators other than ‘information scientists’. There are, for example, the psychologists and sociologists who study people’s behaviour in general; there are those who devise and produce communication channels (from ink-on-paper to the internet); there are those concerned with the symbolism of knowledge representation (linguists); and those concerned with the knowledge content within ‘documents’ (e.g. data mining, geographic information systems, Peter Willett’s chemoinformatics). The totality of activity related to information is today necessarily a multidisciplinary exercise, and the information practitioner, trying to ‘do better’, may draw insights from any of these areas of investigation. His particular concern is with the process as a whole – the interaction between the three elements that leads to people becoming informed.

3. The computer

It has taken some time for the full potential of the digital computer to be realized. It was at first used for numerical ‘computing’, representing numbers in a binary code. But this code could be used to represent any symbol, so the computer came to be seen as a tool for ‘symbol manipulation’ in general. Eventually,
it was shown that any representation – numbers, text, sound, music, images, ‘movies’, ‘data’ of every kind – could be recorded in binary code. How could it be manipulated? As Ada Lovelace said of Charles Babbage’s Analytical Engine, the computer ‘can do whatever we know how to order it to perform’.

Ordering a computer to perform (programming) was a skilled and at first very laborious task. Getting the program instructions and the coded data into the computer (via punched cards) was equally laborious. The space inside for storage and manipulation was at first very limited. Computer ‘printout’ was very clumsy. It would be some time before computers could ‘talk’ to each other via telecommunications. And all this was quite outside the experience of the library/information profession. How could we make use of the computer for our purposes?

This was the challenge that faced, and still faces, the profession. And it has been continually changing. We have been through computer-aided production of abstracts journals; their tapes offered for current awareness search; citation indexes (‘how we learnt to love the Yanks’); fast-response online searching; remote telecommunication access to computers; bibliographic databases with machine-readable records; online search intermediaries and interfaces; OPACs; the internet; the www protocol; an enormous expansion of websites; graphical interface browsers; search engines; and much more.

Ever since the advent of the digital computer, there has been a continual drive to see, as Robert Fairthorne put it in 1961 [3], ‘how far we can go by using ritual in place of understanding’. Robert lived to see the arrival of our current universal digital automaton, the internet. Given an electronic connection, the user can now search an ever-growing slice of the world’s documents from his own desktop. (I use the word ‘documents’ as shorthand for every type of record.) At every step of the way, there is nothing but mechanisms ‘making physical responses to physical signals’, as Fairthorne put it, untouched by human hand or mind. If we want the ‘reality’ of a Gutenberg Bible or a Leonardo painting we have still to go to the trouble of visiting the place that houses them. But reproductions of them, and of live events around the world, can be brought into our homes. The world has been reduced to a vast unified information store, which we can browse from our desks.

Writing further of automation, Fairthorne noted that ‘a clerical system seeks to replace as far as possible semantic activities by manipulations of marked objects according to unambiguous rules. Semantics must enter at output and input at least, else the system would be merely a game, but intermediate operations should involve only the internal semantics of the system; that is, recognition of the marks and what actions they should lead to’. On the internet, semantics at input comprises the intellectual effort that has been put into writing texts, creating images, composing music and so on. Semantics at output comprises the intellectual effort of stating a query that represents what you want and of comprehending the response. The internal semantics are the ‘unambiguous rules’ that the computer has been given in its program.

4. What has the internet become?

At first we thought of the internet as a new way to store and access recorded knowledge. We thought we knew how to deal with it – classify it, like the Yahoo! or BUBL directories. It was an information activity that we were familiar with, but using a new medium. But the internet has become vastly more than that. First, of course, it was from the beginning a new form of person-to-person telecommunication, challenging letter mail and the telephone. Second, it has added a mass of ‘documents’, such as images and detailed ‘raw’ data, with which the information practitioner finds it difficult to deal. Third, it has become a medium for chatter (sorry, social networking) about every topic, public or personal, under the sun. Fourth, and most important, it has been seized upon by commerce, and has become a market place for every kind of consumer goods and services, reputable and disreputable. Information providers have to compete for attention in that market. There are, no doubt, still other functions, present or to come. The internet as a whole is therefore wholly beyond the library/information profession’s experience and control.

The internet – that includes a fair percentage of everything being recorded in the world – has an uncontrolled and ever-growing vocabulary. On the user side, it has opened up the practice of information search
to people of every occupation and of none. Many of their searches do not arise from structured intellectual environments; consequently the semantic relations important to them are unpredictable and variable. There has come about a separation of the ‘visible’ web (that is, visible to search engines) from the ‘deep’ web that lies beyond it on the internet, even though the latter is very much richer in content (BrightPlanet [4]).

Much current recorded knowledge has migrated to the internet (or intranets) – not only public knowledge, but also the private knowledge of industrial firms, government agencies and other institutions. If the internet becomes the primary source of recorded knowledge, and access to it is directly available to the user, in what sense will the information practitioner be an intermediary, ‘between’ knowledge and seeker? In the library, even if he was not acting as a reference librarian or information officer, he controlled and guided access to recorded knowledge by his acquisition policy, his classification, his catalogues, his loan policy. What role can he continue to play?

There is surely still a task for information practitioners to perform – guiding access to recorded knowledge by incorporating our understanding (of the records, and of user needs) into the ritual performed by the computer, to improve its performance in handling information. Recent books (e.g. Lambe [5]; White [6]) emphasize the importance of detailed understanding of the information behaviour of users, and of the types of document used, in designing taxonomies and choosing computer search systems for corporate enterprises.

5. New tasks, new skills

Our problem is that this new task – of giving understanding to the computer – calls for new skills. The information profession believed that it was concerned with organizing recorded knowledge so that people could find their way around it. There was a structure to knowledge, which we sought to embody in classified arrangement, facets, thesaural relations, so that people could navigate within it. ‘The core of arguments [in favour of one knowledge structure or another] was generally, not empirical, but philosophical’, writes Stephen Robertson in his paper in this issue. ‘There was resistance to a strictly functional view of such schemes’. Cleverdon’s Cranfield project, already mentioned, ‘tackled the philosophical divisions in the field head-on’.

The approach to retrieval used by web search engines has been to explore how far ritual could go based strictly on analysis of the occurrence of words in texts, taking into account their frequencies in a document and in the collection, their positions in the document, the document length, and so on, as well as the occurrence of links to other documents. More recently, techniques such as ‘latent semantic analysis’ have been introduced, which seek to deduce semantic links between words based on analysis of their co-occurrence in documents. Since the meanings of the words are ignored, the structure of knowledge plays no part.

The real structure of knowledge is immensely complex. But if people are to navigate a ‘knowledge organization system’ (KOS), they must become familiar with it, it must be relatively limited and simple to understand. Consequently the KOS we construct can only depict a small part of the whole, and the choice of what to depict should, as Stephen Robertson emphasizes, be ‘functional’ – i.e. based on tests as to whether it is useful in manipulating information. The structural tools developed by the profession – hierarchical classification, facets, thesaural relations, topical maps, the predicates used in ontology – will still be of use, provided that they can in each case pass a test of utility.

The information profession will continue to need its skills in acquiring knowledge of users and of recorded knowledge, but needs to develop its understanding of how to incorporate this knowledge into computer ritual. This is much more complex than incorporating it into clerical ritual – as in a static card catalogue – because the computer is dynamic, and can manipulate texts, graphics and data at our command, as well as bibliographic records.

Traditionally, the producers of documents (publishers in the widest sense) have been a profession separate from the information profession. But the computer is still a relatively new environment for both professions. Elizabeth Orna’s idea [7] is a good one: that information designers and information managers should interact and collaborate in the development of computer-based information systems. In this way, documents can be designed from the start to make retrieval and use more effective.
6. Looking ahead

I sense an unease, an uncertainty about the future, in a number of the contributions in this issue. Barry Mahon concludes, ‘It remains to be seen if the current professional bodies recognize the benefits to be gained by embracing the ideas created by the newcomers to the information sector’. Tom Wilson writes that ‘some may already believe that the position has been reached that professional education and research are irrelevant to practice’. David Bawden, speaking of information science theory, suggests, ‘We have not fully worked through the insights of the founders of the discipline, let alone replaced them with entirely new insights’. Speaking of developments in practice, Stella Dextre Clarke feels that ‘the greatest part of the credit […] goes to enhanced information technology rather than fundamental new thinking in knowledge organization’. I took Stephen Robertson’s advice and followed up a recommended reference, and found our greatly missed Karen Sparck Jones urging the need to ‘rethink TREC from the bottom up […] on some principled view of what information management is, should be’. And Jack Meadows concludes that ‘information science activities developed over the past 50 years have triumphed, but information science as a separate entity may be on the wane’.

The uncertainties arise in various contexts, but perhaps an underlying cause in some cases is, as Jack suggests, the apprehension that information science may become ‘submerged’ in the larger field of computer science. Authors more directly associated with computer developments (such as Peter Enser and Peter Willett) are markedly more confident of the future.

Is it possible that we are at a moment of pause or turn in digital information? The expansion of the internet was unprecedented. Search engines dealt with it in the only way possible – scan as much as they could and throw it all into an inverted index. TREC helped to establish a search algorithm that performed as well as any alternative tested, though performance could be improved by relevance feedback (to amend the query terms). Online database search – now with nearly 40 years’ experience – was pushed into the background. Yet, wrote Karen Sparck Jones, the difficulty of obtaining appropriate literature test sets has meant that it is still impossible to make comparisons between natural language approaches and the use of controlled language indexing (the ‘Boolean thesaurus’ model).

Now, the search engines are beginning to experiment with ways to help the searcher amend a query. Attempts are being made to include the deep web in searching. Corporate information systems, much more aware of their problems than a general search engine can be, have been spending much effort on taxonomies. Facet analysis, on a small scale, is making a come-back (Vickery [9]). The Semantic Web, though still very experimental, is demonstrating how much more can be done with a computer if you give it enough knowledge. It seems that we are in for a period of experimental development that may lead to new steps forward in information handling.

References