Then and Now:
Contrasts in the scope of information retrieval

Stella G Dextre Clarke, Vice Chair, ISKO-UK

stella@lukehouse.org

This presentation was delivered as an accompaniment and scene-setter for the Tony Kent Strix Award Annual Memorial Lecture in November 2018. The prestigious Award was inaugurated in 1998 by the Institute of Information Scientists. It is now presented by UKeiG in partnership with the International Society for Knowledge Organisation UK (ISKO UK), the Royal Society of Chemistry Chemical Information and Computer Applications Group (RSC CICAG) and the British Computer Society Information Retrieval Specialist Group (BCS IRSG).

The Tony Kent Strix Award is given in recognition of an outstanding practical innovation or achievement in the field of information retrieval in its widest sense, including search and data mining, for example. This could take the form of an application or service, or an overall appreciation of past achievements from which significant advances have emanated. The award is open to individuals or groups from anywhere in the world.

Stella Dextre Clarke is a past winner and currently the Vice Chair of ISKO-UK (International Society for Knowledge Organization).

****

In considering how the scope of information retrieval (IR) may have evolved over the years, this article has twin objectives. Firstly, it should provide some context for those who never had the chance to meet Tony Kent and may wonder why we still honour his leadership and achievement. Secondly it responds to my personal curiosity about the meaning of "Information Retrieval", after a reviewer queried my use of the term in a recent article.

I should explain that I had used the term quite broadly, to include all the steps involved in any kind of searching for information, whether automated or manual. I then applied it more specifically to the context of thesaurus use. My reviewer thought this would not be understood, because, "research in IR has largely migrated to computer science and the term seems to have changed meaning in the direction of search engines." To explore whether/how the meaning of the term has changed, the first part of my presentation compared a modern definition with others from the early days, and especially from the time when the IR pioneers were inspired by Tony Kent. In the second part, I listed the principal IR achievements of the past winners of the Strix Award, looking for any trends that might reveal an evolution in the scope of IR.

Definitions then and now

Let's start with how the term is understood today, and Wikipedia is the obvious place to look:

> *"Information retrieval (IR) is the activity of obtaining information system resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for metadata that describe data, and for databases of texts, images or sounds."*

Arguably that definition can be interpreted to include manual processes, although plainly it will mostly be applied to computer-driven processes. An equally broad but more authoritative definition can be found in the current international standard ISO/IEC 2382:2015, "Information technology — Vocabulary":

> *"Actions, methods, and procedures for obtaining information on a given subject from stored data."*

But now let's go back to the post-war period, long before the era of the personal computer. Calvin Mooers is usually credited with coining the term, and in 1951 he wrote:

> *"Information retrieval is the name for the process or method whereby a prospective user of information is able to convert [his or her] need for information into an actual list of citations to documents in storage containing information useful to [him or her.] Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, or machines that are employed to carry out the operation."*

Understandably there is no mention here of a computer, but the context of Mooers' work was the "mechanical organization of knowledge", in particular a technique he called "Zatocoding". "Coding" was not a synonym for "programming" but referred to a way of expressing descriptors in a short, coded form that could be applied to a card-based system. His examples ranged from a small deck of machine-sortable edge-notched cards to the immense stack of 80-column IBM cards needed for a collection the size of the Library of Congress (5 million documents in those days).

Scientists from the Royal Society of London were keenly interested in IR, which they saw as dependent on classification. This led to establishment in 1952 of a Classification Research Group (CRG), which in 1957 published a Memorandum entitled "The Need for a Faceted Classification as the Basis of All Methods of Information Retrieval." While recognising four distinct "mechanisms" for IR, namely indexing, classification, automatic selection (e.g. Mooers' Zatocoding) and co-ordinate indexing (e.g. Mortimer Taube's Uniterm method), the memorandum argued that a standard faceted classification should be the basis for all of these.

A decade later scientists and engineers were still pressing for R&D in IR, with much the same scope in mind. In the *Aslib Handbook* of 1967, John Sharp wrote "information retrieval is taken to cover all the techniques, conventional and non-conventional, which

are used to provide for the recovery from a store of documents of those items which are relevant to a stated information need". Like Mooers and the CRG, his definition included whatever classification, indexing or other processes might be needed; and all of them assumed the aim of IR was to identify relevant documents in some kind of storage.

At around the same time Tony Kent had agreed to head a research unit established at Nottingham University with the backing of the British Chemical Society. Their aim was to investigate potential uses of the machine-readable tapes used in publishing *Chemical Abstracts*. The tapes were a by-product of the automated production process for the printed journal, not designed with IR in mind. But Tony was not just a zoologist and keen birdwatcher; he was already fascinated by the potential of computers to extract information from his ornithological records and determined to explore what more could be achieved with structured text data.

IR expectations in the sixties and seventies

To appreciate the boldness of Tony Kent's venture, we need to contrast his pioneering spirit with the mindset of IR front-runners at that time. It is true that Vannevar Bush had anticipated key IT functionality in 1945, with his hyperlinked "Memex" machine. Arguably Paul Otlet had too, in his Mundaneum vision in the years from about 1910. And the computer was not unheard of. But still by 1970, computer use was way beyond the budget or floor space of the typical library or information centre. Use of a computer for IR purposes was beyond the wildest dreams of most researchers. Tony Kent's vision seemed like pie in the sky.

Predominant in the 1960s (and beyond) was an expectation that thorough IR must always somehow depend on classification. Even if computer-generated KWIC (KeyWords In Context) indexes – and variants such as KWOC (KeyWords Out of Context) – were proving their worth for current awareness services, their weakness was seen to be reliance on *words* rather than *concepts*. Classification was recognised as the key technique to analyse the *subject* of a document rather than the *terms* to be found in it.

At the same time, several alternative or supporting technologies were on the up, among them:

- Microfilm and microfiche, allowing a larger collection to be stored in the same space;
- A huge variety of card systems – edge-notched cards, feature cards, item cards, aperture cards, 80-column cards, peekaboo cards, optical coincidence cards, machine-sortable cards, even ordinary catalogue cards – for the indexes and sometimes abstracts of the items in the collection;
- Computer-generated KWIC and KWOC indexes, even SLIC (Selective Listing In Combination) indexes, carrying SDI (Selective Dissemination of Information) services from the likes of the American Chemical Society and the National Library of Medicine.

And there was much research into use of computers for analysis, classification and indexing of documents, including machine translation. Techniques based on sorting delivered early successes.

From the literature of the 1960s, here are a few verdicts from researchers into the IR potential of computers:

> *"So far as indexing and searching go…good 'manual' systems are still every bit as effective in the vast majority of cases, and very much cheaper".*
>
> Jack Mills (1963)

> *"The costing of computer-based systems seems…to be almost fatuous [on grounds firstly of effectiveness and secondly of the high prices paid]".*
>
> John Sharp (1967)

> *"It is here [automated SDI] that mechanization offers possibilities".*
>
> Wilfred Ashworth (1967)

I hope the above quotations illustrate the scepticism that confronted Tony Kent when he took on his Nottingham appointment, not to mention the unknown quantity of any text processing vision or methodology. Undaunted, he went on to launch the commercially successful UK Chemical Information Service (UKCIS), and his leadership inspired a great many others to follow. A lot more about Tony's subsequent achievements can be found in a booklet about the Tony Kent Strix Award, assembled by the organising committee and downloadable now from the Strix Award web page. In Section 5 of that booklet, Jan Wyllie quotes two sceptical, or at least cautionary, remarks from the great man himself:

> *"I refuse to believe that knowledge can be inferred from any conceivable software system".*
>
> - from Trend Monitor Reports, July 1991

> *"Real literacy (as opposed to computer literacy) is a necessary prerequisite for the effective use of information, and…computer technology can only, at best, provide gadgets that reduce drudgery".*
>
> – Ibid., December 1989

Moving forward

Since those early days five decades of research and technology progress have transformed the scene and brought IR from the wish-list of the scientist to the fingertips of the general public. And has the meaning or scope of "information retrieval" changed in that time? Maybe some clues can be found in the list of topics of past winners of the Strix Award – (see Table 1). Or maybe not. No clear trend stands out for me. In recent years one interesting feature is an emphasis on the human side of things. Plainly the nominators and judges have been impressed by leadership in support of the user, or to encourage communities of IR students, researchers and developers.

Table 1: Past Award winners, and the IR achievements for which they are noted

Year/Winner	Principal achievements
2018 Pia Borlund	IR user studies, evaluations and test design, especially the Interactive Information Retrieval (IIR) evaluation model.
2017 Maarten De Rijke	Computational methods for analysing, understanding and enabling effective human interaction with information sources.
2016 Maristella Agosti	IR community leadership, as well as research in hypertext, digital libraries, evaluation methodology and more.
2015 Peter Ingwersen	Theoretical understanding of IR, applying this notably to integration of IR and human information seeking processes.
2014 Susan Dumais	Research at the intersection of Human-Computer Interaction (HCI) and IR, such as co-invention of Latent Semantic Analysis and Indexing (LSI).
2013 W Bruce Croft	Clustering, passage retrieval, sentence retrieval and distributed search, ranking functions, language modelling, and more. Croft was a distinguished IR all-rounder.
2012 Doug Cutting and David Hawking	The Award was shared between Cutting, who developed Lucene and Hadoop software; and Hawking, the coordinator of two tracks of the Text REtrieval Conference (TREC) who also developed enterprise search software.
2011 Alan Smeaton	Techniques for Natural Language Processing (NLP) in text as well as for indexing and retrieval of non-text data.
2010 Michael Lynch	Variety generation, applied firstly to chemical substructure searching and then more generally, e.g. to databases of chemical reactions.
2009 Carol Ann Peters	Leadership and sustained development work on the Cross Language Evaluation Forum (CLEF).
2008 Kalervo Jarvelin	NLP method evaluation, ontology-based query expansion and relevance feedback, cross-language IR (CLIR) methods/evaluation and IR evaluation metrics.
2007 Mats G. Lindquist	Digital library work, including a lead role in the Paralog IR software.
2006 Stella Dextre Clarke	Development of GCL/IPSV classifications for the UK public sector, plus work on British and International thesaurus/interoperability standards.
2005 Jack Mills	Research on faceted classification, the Cranfield IR project, and revision of the Bliss Classification scheme.
2004 Keith van Rijsbergen	Theoretical modelling of IR systems.

2003 Herbert van Sompel	Development of the Open Archives Initiative (OAI) and standards such as OpenURL, Object Reuse and Exchange, and the OAI Protocol for Metadata Harvesting.
2002 Malcolm Jones	Research leading to implementation of the web-based *Encore!* union catalogue of musical performance sets, and development of the International Standard Music Number (ISMN) standard.
2001 Peter Willett	R&D in chemoinformatics and many other standard capabilities of IR software.
2000 Martin Porter	Developing the Porter stemming algorithm and later the IR software of Muscat and derived commercial products.
1999 Donna Harman	Leadership of the TREC.
1998 Stephen Robertson	Probabilistic methods of IR, notably the BM25 ranking algorithm, coupled with interface design and other aspects first demonstrated in the OKAPI software. Contributions also to the TREC.

What really has changed is the social and technological context in which IR is applied, leading to a huge expansion in scope, (see Table 2). The first challenge for information professionals in the pre-internet days was often getting hold of literature from remote places. In the 1960s an in-house collection was indispensable, and much effort was put into selection of its content, coding and microfilm, etc., all with the object of keeping things small, manageable and affordable. Nowadays size is not an issue, as resources from all over the world can be accessed via electronic networks. The ubiquity of computers and smartphones has brought a need for IR to almost everybody. With a computer built into the car engine, the phone, even the oven and the refrigerator, the scope for IR has expanded almost beyond recognition. As the Strix Award list illustrates, IR has applications in all fields, from music to chemistry and many, many more, without linguistic limits.

Table 2: Contrasts in the scope of Information Retrieval

Then (1960s)	**Now (2019)**
Applications of limited size	The only size limit is in your imagination
In collections, especially libraries but also some bibliographic databases (on cards or on magnetic tape)	The same, plus virtual collections, networked resources, non-text media, multilingual content, unstructured data, intranets, PC content, and more
Classification, indexing, sorting, citation indexes, hyperlinks in theory	The same, plus ranking, stemming, filtering, LSI, clustering, linked data, etc.
Computer use a rare luxury, value disputed, reliant on batch processes	Computer use the norm, mostly online and interactive. And computers are *everywhere*
Led by scientists, engineers, professional societies, librarians	Consigned to the IT department? Or is it *everyone's* business?

Conclusions?

Concerning the definition of information retrieval, the early ones were not based around computer science, even less around search engines, and this has not changed in current definitions. That said, computer use is the norm for almost every task nowadays, and there is no denying the prevalence of search engines. I suggest that computer science will continue to fuel advances in IR, but not in an exclusive way. There is a continuing opportunity and a need for imaginative IR enthusiasts, in the mould of Tony Kent, from all fields and any walk of life.

I suggest also that in many contexts the activities of metadata preparation, classification, indexing, etc. are considered valid components of IR, as they were in the early days. Data storage too is still within the field, though not essential to a particular IR task.

The assumption that it was enough to retrieve relevant *documents* has certainly moved on, in those systems which seek to pinpoint the relevant paragraph, word, phrase or character string.

In the modern context the Wikipedia definition "… the activity of obtaining information system resources relevant to an information need from a collection of information resources" could do with an update, as the stipulation of a collection seems debatable. When we search using Google, arguably there's a "virtual" collection, but "nebulous", "arbitrary" or even "non-existent" might describe it better. And as to basing IR on an "information need", what about serendipitous finds from surfing the Internet – a retrieval activity or not?

Most of the time we function intuitively without careful definitions, and no doubt you have your own view of what IR should cover. I'm sticking with my broad interpretation of the scope!

And where next for IR?

Just a few of the expanding opportunities for IR include the following:

- Finding a particular nugget of information amongst the deluge - still a challenge despite (or perhaps because of) the oceans of information available to us all.
- IR still has a long way to go with multimedia collections, especially audio resources
- The Internet of Things will offer unlimited scope.

While preparing this presentation, I tried to let my mind run free over the developments I'd really like to see. The first occurred to me when putting together the reference list you'll find below. I had already assembled the various quotations mentioned above, by looking online, and in the various directories and databases on my PC, and from the runs of journals, textbooks and anthologies that line my study. Frustratingly, I had not recorded where I found each of them. What if some future IR capability would let me run a search over all those resources, printed and electronic, at one fell swoop? That may sound

ridiculous to the point of unthinkable, but much of what we take for granted now was unthinkable in the sixties.

What more could I wish for? Sometimes when I delve into my old files, I'm amazed to read things written by myself that once upon a time I must have known thoroughly. And I am not the only one! Maybe you too wish you could easily retrieve all your long-lost memories? Could it be that the future will see some kind of convergence between IR and the research into Alzheimer's disease, enabling all of us to function more effectively?

Just as the pioneers in the sixties could not have foreseen how IR would expand into the 21st century, so our predictions for the next fifty years are unlikely to be accurate. But there's hope, and there's scope, for many exciting IR successes to come.

References

Ashworth, Wilfred. "A Review of Mechanical Aids in Library Work." *Handbook of Special Librarianship and Information Work*. 3rd ed., Editor Wilfred Ashworth, 524-53. London: Aslib, 1967.

Bush, Vannevar. "As We May Think." *Atlantic Monthly* 176 (1945): 101-8.

Classification Research Group. "The Need for a Faceted Classification as the Basis of All Methods of Information Retrieval" *Proceedings of the International Study Conference on Classification for Information Retrieval*. London: Aslib, 1957.

Dextre Clarke, Stella G. "Thesaurus (for information retrieval)." Published (2017) in the *ISKO Encyclopedia of Knowledge Organization*.

Mills, Jack. "Information retrieval: a revolt against conventional systems?" *Aslib Proceedings*, Vol 16, No 2, Feb 1964, pp 48-63

Mooers, Calvin. "Zatocoding applied to mechanical organization of knowledge." *American Documentation,* 2 (1951), 20-32

Sharp, J. R. "Information Retrieval." *Handbook of Special Librarianship and Information Work*. 3rd ed., Editor Wilfred Ashworth, 141-232. London: Aslib, 1967.

Sparck Jones, Karen, and Peter Willett. *Readings in Information Retrieval*. San Francisco, USA: Morgan Kaufmann, 1997.