A history of enterprise search 1938-2022

A history of enterprise search 1938-2022

MARTIN WHITE

Contents

Part I. Main Body

Author biography

Martin White is Managing Director of Intranet Focus Ltd, a consulting business based in the UK. His work on intranet implementation indicated that many organisations were struggling to achieve a high level of satisfaction with search applications and since 2008 most of his client projects have focused on enterprise search implementation.

Martin is a Fellow of the British Computer Society and of the Royal Society of Chemistry, an Honorary Fellow of the Chartered Institute of Library and Information Professionals, and a member of the Association for Computing Machinery (USA). He has been a Visiting Professor at the Information School, University of Sheffield, since 2003, where he lectures on information management and information retrieval. He is the author of ten books on various aspects of information management technology.

Preface

As you walk up Walton Street in Oxford the road bears slightly to the left and a large 19th century building comes into view. It is not an Oxford college but the headquarters of the Oxford University Press. OUP is the largest university press in the world, and can date its origins back to around 1480. In 1983 I arrived at this building carrying a Texas Silent 700 terminal. This used thermal ink printer technology and had two rubber ears on the top into which a telephone handset could be inserted to link the printer into the public telephone network. A decade earlier I had used the same technology to use the first computer-based search services developed by the Lockheed Corporation and System Development Corporation.

I was heading up early attempts by Reed Publishing to develop electronically published products and services, notably airline flight timetables. Reed owned International Computaprint Corporation, based in Fort Washington, PA, which specialised in keyboarding and printing telephone directories. Reed had been working with IBM and the University of Waterloo, Canada on the New Oxford English Dictionary (NOED) project, which was to create a digital version of the Oxford English Dictionary. The proof of concept was to digitise the one of the Supplements to the First Edition, starting at the letter S. The digitisation and indexing had now been completed and I, together with Hans Nickel, the founder and CEO of ICC, were to demonstrate what we had achieved to the NOED project team, led by Tim Benbow and Edmund Weiner. Many of the team of lexicographers were sceptical of the value of the project, and there was a mixture of expectation and disinterest around the table.

The OED seeks not only to provide a definitive definition of a word, but also the origins of when the word was first used, with examples of subsequent use which may have modified the definition. All these examples were contained on around four million slips of paper. With the terminal we set up a connection (at 300 baud) to the computer in Fort Washington. I can still remember the first question, which came from one of the more sceptical lexicographers, who wanted to know how many words in the OED originated in the Times newspaper. Because all the text had been marked up in Standard Generalised MarkUp language (a forerunner of XML) we could identify the source, and not only provide a count but print out (albeit very slowly) all the examples. There was a short period of silence and then these distinguished scholars suddenly realised the potential of information retrieval. They also recognised that it was not going to put them out of a job but enable them to improve the value of the product. Many more queries were undertaken and the session only came to an end when we ran out of supplies of thermal paper.

The NOED project was an enormous success, not only for the OUP but also for Dr Gaston Gonnet and his team at University of Waterloo. This team became the nucleus of Open Text Corporation. IBM used the knowledge gained from the project in the development of its search technology as the OED files provided a rich source of syntax information to help with query development.

For me it was a day of discovery about the power of search to discover new relationships between items of information. I learned three important lessons from this project. The first of these was the value of metadata structure in searching. Because of the way that the individual elements of the entries had been marked up in SGML it was easy to search for words that had first been used by Charles Dickens after his return from his first visit to the United States in 1842. The second lesson was gained in listening to the members of the project team from IBM and the University of Waterloo as they talked about the importance of computers being able to understand the structure of sentences, work that would lead to the development of semantic search technologies. The third lesson was in understanding the impact that search could have on organisational processes and outputs.

Introduction

I am in the fortunate position of knowing exactly when and where I was introduced to the use of computers to search for information. The date was 23 February 1976 and the location was the Institution of Electrical Engineers offices in Savoy Place, London. The occasion was a presentation by the UK Department of Trade and Industry of a UK link to the RECON service of the European Space Agency, based in Frascati, Italy. Remote access time-share research services had been available in the USA for over a decade but access to the services from the UK was technically difficult and very expensive. The definitive book on the development of online information services from 1963-1976 (Bourne & Hahn, 2003) runs to over 500 pages on just this fairly narrow but very important period of search technology development.

The development of computer hardware and software since the 1950s has been documented in the IEEE Annals of the History of Computing and in a number of books, notably A *new history of modern computing* published in 2021 (Haigh & Ceruzzi, 2021). There seems to have been no history of enterprise search which covers both the development of the technology and also its commercial exploitation. This book is an attempt to provide an overview of enterprise search, starting with the adoption of punched-card systems in the late 1930s and ending with the arrival of AI/ML technology in the 2020s.

References

Bourne, C.P. & Hahn, T.B. (2003). A *history of online information services from* 1963-1976. MIT Press

Haigh, T. & Ceruzzi, P.E. (2021). A *new history of modern computing.* MIT Press.

1. 1938 -1948 Punched cards as the genesis of enterprise searching

The choice of the year 1938 is somewhat arbitrary. From the mid-1930s onwards in the USA in particular the use of punched cards to enable collections of information to be sorted was gradually being adopted. Punched cards were initially developed by Hollerith to help the US Census Bureau process the 1890 Census, taking as a model the Jacquard loom. This loom had been invented by Joseph Marie Jacquard in 1804, using punched cards linked together to create complex patterns.

The adoption of punched cards to manage book and report catalogues started to be more widely adopted in the late 1930s but still on a small scale. Moving into the 1940s, and unbeknown to the library community, punched cards were being used on an industrial scale by the code-breaking teams at Bletchley Park (UK) to manage the analysis of decoded messages in order to create operational intelligence about the movement of enemy military units and personnel. Towards the end of WW2 Bletchley Park was processing two million cards a week. The techniques used to manage these cards remained secret until the 1970s. However, the initial outcome was the availability of very robust card tabulators that were on show at the 1948 Royal Society Conference without any indication of their origin.

During WW2, the rapid growth in research in the USA in particular (especially in chemical synthesis) led to a very substantial growth in published research after the war had ended. Chemical Abstracts, the central abstracting publication for the field worldwide, shows 33,672 abstracts published annually in 1945; by 1950 it had reached 59,098; and by 1955, 86,322 (57% and 68%

growth rates respectively in the five-year periods). Much of this growth was in organic chemistry, where the development of infrared spectroscopy in particular led to important advances in determining the structure of organic compounds and then assessing the activity of pharmaceutically active compounds to their chemical structure.

The problem that chemists had faced for many years was that it was possible for a given chemical entity to have a number of text descriptions, leading to a significant amount of confusion.

For example, the chemical formula CSCl4 could be described as

- Perchloromethyl mecapatan
- Thiocarbonyl tetrachloride
- Trichloromethyl sulphur chloride
- Tetrachloromethyl thiol
- Trichloromethyl sulfenyl chloride

To make matters worse there were British, French, German and American naming conventions.

A solution to this problem was developed by the British chemist George Malcolm Dyson (1902-1978) who developed a linear alphanumeric code that was unique to each structure.

The first announcement of what would become known as the Dyson Notation was a letter by Dyson dated 24 June 1944 and published in *Nature* on 22 July 1944. In the letter he mentions that he would be publishing a book on the systematic notation that he was developing. He stated the objective as establishing a database (though he did not use this term) of codes, each of which represented the structure of a unique chemical entity. The notation was based around determining and then supplementing the longest carbon chain.

The first public presentation by Dyson of his notation for organic compounds was at a meeting of the Royal Institute of Chemistry in 1946. The Institute was so impressed it circulated a copy of his lecture to its members. The first edition of his book A New Notation and Enumeration System for Organic Compounds was published by Longmans in 1947. Then on 3 February 1948 he gave a lecture to the British Society for International Bibliography that was reprinted in the inaugural issue of *Aslib Proceedings* (Dyson 1949) along with the discussion which followed his presentation. A second edition of his book was published in 1949. The major change between the editions is a final chapter on the potential of punched cards for managing chemical information.

In the development of his notation Dyson had built up a friendship with James Perry, a highly respected chemist working in the Library at MIT. Both could see the potential to manage chemical information using punched cards. This led to Dyson and Perry meeting with Thomas Watson, the President of IBM, though sources differ if this meeting took place in 1948 or 1949. Watson was impressed with their vision and arranged for H.P. (Pete) Luhn to work with them on developing punched card devices for information retrieval.

By now the benefits of using punched cards by major pharmaceutical companies in the USA and the UK as a means of searching through collections of reports was becoming very evident, and the processes they used could certainly be described as enterprise searching. It was the combination of these processes and the advent of computers that could transform the selection process from a mechanical tabulator to a digital machine that formed the basis for the evolution of enterprise search as we see it today.

A full account of the adoption and development of punched card systems (often referred to at the time as 'mechanical indexing') and the transition to digital storage and search has been prepared by Robert Williams (Williams 2002) who was in the forefront of this work in the USA and writes from personal experience of the pioneers.

References

Dyson, G.M. (1949). International chemical abstracts and the new notation for organic chemistry, *Aslib Proc.*, **1** (1) 5-21

Williams, R.V. (2002). The Use of Punched Cards in US Libraries and Documentation Centers, 1936-1965. *IEEE Annals of the History of Computing*, 24(2), 16-33. https://ieeexplore.ieee.org/document/1010067

2. 1949 – 1959 The dawn of computers

1948 was an auspicious year in the development of both scientific information management and the use of computers to search text files. The Royal Society Scientific Information Conference identified the challenges that lay ahead in managing the flow of scientific information; challenges that arguably we have not solved. The earliest research into how computers might help was undertaken by Philip Bagley (Bagley 1951) as part of a Masters project at MIT. His thesis was entitled Electronic Digital Machines for High-Speed Information Searching. He set out the basic principles of 'information searching' and wrote a program for the Whirlwind computer at MIT.

Following graduation, Bagley was employed at MIT Lincoln Laboratory, and then at MITRE Corporation, where he worked on the SAGE air defense system. In 1964 he moved to the Philadelphia area to enter graduate school in Computer and Information Science at the University of Pennsylvania.

He submitted his PhD dissertation in 1969, in which he coined the now widely familiar term 'metadata' but the thesis was not accepted, and published only as a report under contract with the Air Force Office of Scientific Research, entitled, Extension of Programming Language Concepts.

By June 1952 there was enough interest in the subject at a number of research centres across the USA to hold a Symposium for Machine Techniques for Information Selection at MIT. One of the speakers at the Symposium was Hans Peter Luhn, at that time working on punched-card retrieval systems for IBM. Luhn would turn out to be hugely influential in information retrieval and his hash algorithm (which he developed in the late 1950s) remains in use to this day.

Another very influential person was Eugene Garfield, who in 1955 published a paper in *Science* about the value of citation analysis. (Garfield 1955). From this approach Garfield launched his Institute for Scientific Information to commercialise citation analysis. His insight also became one of the innovations incorporated into Google at the outset in the 1990s, but that is another story. Of more immediate interest is a paper by Allen Kent and his colleagues at the Battelle Memorial Institute, Ohio. In this paper (Kent et al 1955) the concepts of 'recall' and 'pertinency' are proposed as metrics for a search application.

There were two further important conferences in the 1950s.The first was the International Study Conference on Classification for Information Retrieval, held in Dorking, UK in 1957. This was the first opportunity for UK and US research teams to exchange ideas and research on information retrieval. The USA may have had a technology lead, but the UK was held in high regard for research and implementation of classification and index frameworks.

A year later an International Conference on Scientific Information was held in Washington D.C. to take note of developments since the 1948 Royal Society conference and much of the discussion was about information retrieval. The papers make for some fascinating reading. By 1958 Dow Chemicals was evaluating how computer-based systems could be used to manage in-house documentation.

The chemistry community has some special information retrieval challenges (such as searching chemical structures) and has always been in the vanguard of search development. It was at an American Chemical Society meeting in Miami in 1957 that Luhn gave a paper on A statistical approach to mechanized encoding and searching of literary information (Luhn 1957) in which (in effect) he set out the constituent elements of a search application.

The following year Luhn published a paper on his work at IBM (Luhn 1958) in which in which (according to the abstract):

"Excerpts of technical papers and magazine articles that serve the purposes of conventional abstracts have been created entirely by automatic means. In the exploratory research described, the complete text of an article in machine-readable form is scanned by an IBM 704 data-processing machine and analyzed in accordance with a standard program. Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the 'auto-abstract'."

This was indeed a visionary approach. Luhn also proposed that the frequency of word occurrence in an article furnished a useful measurement of word significance. This is the origin of the now familiar term frequency – inverse document frequency model although it was not until 1972 that Karen Spärck-Jones developed a rigorous statistical basis for TF.IDF.

In 1959 Maron and Kuhns wrote a seminal paper entitled On relevance, probabilistic indexing and information retrieval (Maron and Kuhns 1960) in which in which they defined 'relevance' (to replace 'pertinency' and the use of 'probabilistic indexing' to allow a computing machine, given a request for information, to make a statistical inference and derive a number (which they called the 'relevance number') for each document. They suggested that this could be a measure of the probability that the document will satisfy the given request. The result of a search would then be an ordered list of those documents which satisfy the request, ranked according to their probable relevance. The achievement of high levels of relevance has since become the Holy Grail of enterprise search.

The importance of the paper is that Maron and Kuhns then evaluated their proposal through a manual (rather than computer-based) trial, so setting out not only the fundamental principle of determining the probability that a document was relevant but the importance of system evaluation. Fifty years later Maron published a short account (Maron 2007) of the background to this paper in which he provides a fascinating insight into how he and Kuhns developed this principle.

The transition from cards to computers is described in detail by both Harman (Harman 2019) and Robertson (Robertson 1994) A number of papers on the early history of the adoption of computers into the production of Chemical Abstracts were given at a conference held in 2014 on the Future of the History of Chemical Information.

Although Maron and Kuhns had shown that a probabilistic approach was superior to a Boolean approach, virtually all of what might be seen as the first generation of commercial search applications used Boolean logic because the challenge of calculating a 'relevance number' had yet to be solved. It is of note that Maron was at the RAND Corporation which had set up System Development Corporation (SDC) as a subsidiary. RAND spun off the group in 1957 as a non-profit organisation that provided expertise for the United States military in the design, integration, and testing of large, complex, computer-controlled systems. SDC became a for-profit corporation in 1969 and began to offer its services to all organisations rather than only to the American military. It played an important role in search development. Another important development in 1959 was the establishment of the Augmentation Research Center at Stanford Research Institute under the direction of Doug Engelbart.

By the end of the 1950s almost all the core elements were in place, including understanding the required modularity of the search process, the benefits of a probabilistic view of document retrieval, the concepts of precision, recall and relevance, and the value of testing and evaluation. What was needed now was computing power to provide an acceptable level of responsiveness when searching large collections of documents.

References

Bagley, P.R. (1951). Electronic digital machines for high-speed information searching. MIT Press. http://hdl.handle.net/1721.1/12185

Garfield, E. (1955). Citation indexes for science: a new dimension in documentation through association of ideas. *Science*, 122(3159), 108-11.

Kent, A., Berry, M.M., Luehrs Jr., F.U., & Perry, J.W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation*, 6(2), 93-101. https://onlinelibrary.wiley.com/doi/10.1002/asi.5090060209

Luhn, H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal*. October 1957

Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM Journal*. April. https://ieeexplore.ieee.org/document/5392672

Maron, M. & Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the* ACM. July. https://dl.acm.org/doi/10.1145/321033.321035

Maron, M.E. (2007). An historical note on the origins of probabilistic indexing. *Information Processing and Management*, 44, 971-972.

Harman, D. (2019). Information retrieval: the early years. Foundations and Trends in Information Retrieval, 13(5), 425-577. http://dx.doi.org/10.1561/1500000065

Robertson, S.E. (1994). Computer retrieval as seen through the pages of the Journal of Documentation. In B.C. Vickery (Ed.) Fifty years of information progress. (118-146). Aslib.

3. 1960-1969 The pioneers

Condensing the immense amount of progress made in the 1960s is not easy and so this is a very selective perspective. As far as algorithm developments were concerned Bourne and Ford published a paper on stemming in 1961 (Bourne and Ford 1961), Damerau (Damerau 1964) reported on approaches to solve misspellings and Rocchio and Salton considered how best to optimise the performance of retrieval systems Roccio and Salton 1965). This was one of the first outcomes of the SMART project, initially at Harvard and then at Cornell, that will figure significantly in the history of the 1970s. Many of the developments of the period were reported in a new Information Retrieval section of ACM *Communications* from March 1964. A year earlier *Information Storage and Retrieval* was launched as a peer-reviewed journal, changing its name to *Information Processing and Management* in 1975.

Another initiative that started in the 1960s and lasted into the 1970s was ground-breaking work by Cyril Cleverdon, the librarian of the Cranfield Institute of Technology, UK on the comparative efficiency of indexing systems. It was funded by the US National Science Foundation. I had the good fortune to meet Cyril early in my career and his encouragement of my career choice was along the lines of "You will never be out of a job". How right he was!

In the 1960s advances in computer technology resulted in some very technical progress in search development in terms of both research and the availability of commercial services. IBM released the 7090 range in late 1959 and the much more powerful 360 range in 1965. In parallel the technology to provide remote shared access to large computer centres was developed, with J.C.R. Licklider as the early innovator, leading directly to the Internet. At this point in the history of search a strictly chronological approach is not of value, and instead it is important to be aware of a number of major projects, several of which led to commercial online services becoming available from 1965 onwards.

Arguably the first ever enterprise/internal search service was set up in 1965 at the Cox Coronary Heart Institute in Kettering, Ohio by G. Douglas Talbott. I would cite this as enterprise search because the application indexed content that the Institute was publishing in a quarterly internal publication

In terms of the impact on the underlying algorithms of search, the work at System Development Corporation in the early part of the decade is of particular importance. Synthex was led by Robert Simmons with the objective of developing a system that could read and understand text, answer questions and compose an answer in readable English. The name was chosen as a tribute to the Memex concept of Vannevar Bush from 1945. There was a related ProtoSynthex project. One outcome of these projects was TEXTIR, an online search system developed for the Los Angeles Police Department in 1964 that could accept questions in natural language. Further development enabled it to incorporate synonyms into a search formulation and offer search term weighting. In parallel Hal Borko (Borko 1964) was developing BOLD with a focus on the automatic classification of the text in documents. Yet another project was COLEX, the aim of which was to advance the development of time-sharing services to provide online access to bibliographic databases.

These projects gave SDC the ability to launch the ORBIT online search service in 1967, a commercial service for information professionals and researchers which enabled them to search through large databases of abstracts of research literature. The project was led by Carlos Cuadra. Just a few months earlier the Information Sciences Group at the Lockheed Palo Alto Research Laboratories, led by Roger Summit, had launched the DIALOG online search service. The focus of this group was more towards scaling up online services and user interface development and one of its innovations was the display of set numbers at each stage of a query, a forerunner of facet hit numbers in current search applications. However probably the first public demonstration of computer-based information retrieval was at the 1964 World Fair with the LIBRARY/USA demonstration.

Other major centres of information retrieval science and application development in the 1960s included the work at Harvard and then Cornell University led by Gerard Salton, though this did not come to fruition until the early 1970s. Probably the most innovative was the work of Donald Hillman at Lehigh University on searching the full text of documents (the LEADER project) but mention should also be made of the SPIRES project at Stanford University (which remains one of the pre-eminent centres of information retrieval to this day) and TIP at MIT's Lincoln Laboratories. IBM was also very much involved in retrieval research on a global basis and research into the use of computer applications for law research had been initiated. These and many other projects are described in detail by Bourne and Hahn in The History of Online Services 1963-1976 [v] and in addition there is an excellent paper by Hahn (Hahn 1996) based on the research for their book.

The importance of these online services to enterprise search is that they addressed the issues of scaling up the concepts developed in the 1950s and started to pay attention to user satisfaction, the user interface and user support. Probably the first user assessment of an online service was carried out in 1969 by Timbie and Coombs (Timbie and Coombs 1969). It was not until the early 1970s that these services were available in Europe and indeed globally, a problem primarily of low network capacity and very high network access costs. The launch of these services also set a standard for the search experience for a generation of information professionals and researchers that was not challenged until the arrival of Alta Vista and then Google 30 years later. These online services showed that research services could be delivered on demand at the desktop. The next decade was primarily about improving search result relevance and performance.

References

Borko, H. (1964). Research in automatic generation of classification systems. AFIPS ’64 (Spring): Proceedings of the April 21-23, 1964, spring joint computer conference.

Bourne, C.P. & Ford, D.R. (1961). A study of methods for systematically abbreviating English words and names. *Journal of the* ACM, 8(4), 538-552. https://doi.org/10.1145/321088.321094

Bourne, C.P. (2003). A *History of Online Information Services*, 1963-1976. MIT Press.

Damerau, F.J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3), 171-276. https://doi.org/10.1145/363958.363994

Hahn, T.B. (1996). Pioneers of the online age. Information Processing & Management, 32(1) 33-48. https://www.sciencedirect.com/science/article/abs/pii/030645739500048L?via%3Dihub

Roccio, J.J. & Salton, G. (1965). Information search optimization and interactive retrieval techniques. AFIPS ’65 (Fall, part I): Proceedings of the November 30-December 1, 1965, fall joint computer conference, part I. November, 293-305. https://dl.acm.org/doi/10.1145/1463891.1463926

Timbie, M. & Coombs, D. (1969). An interactive information retrieval system – case studies on the use of DIALOG to search the ERIC document file. ERIC Clearinghouse on Educational Media and Technology at the Institute for Communication Research, Stanford University.

4. 1970-1979 Enterprise search emerges

In the 1970s products emerged which are clearly the antecedents of what we would regard as enterprise search applications. From here on in the focus on academic research in this history will be significantly less, not because less research is being carried out but because it is well documented in a range of books. In particular each chapter of Introduction to Information Retrieval by Manning, Raghavan and Schutze (Manning, Raghavan and Schutze 2008) has an annotated bibliography and can be downloaded as a pdf. However, there are three academics that deserve mention. The first of these is Gerard Salton. He developed the SMART software application as a 'test bed' at Harvard University and took it with him to Cornell University where he stayed for the rest of his career. Salton developed the cosine vector space model (VSM) to compare the relevance of a group of search results. The evolution of this model took place over a number of years and David Durbin has tried to unravel the way in which it developed, providing a good bibliography.

Karen Spärck Jones worked in a number of departments at Cambridge University from the time of her PhD in 1964. A profile of her work whilst at Cambridge links to papers describing her research, all of which has had a major impact on information retrieval. Her overview of information retrieval research (Spärck Jones 2006) is essential reading. The third person is Stephen Robertson, a research colleague of Karen Spärck Jones, who went on to work at the Microsoft Research Laboratories in Cambridge. His work has extended from the mid-1970s until quite recently, the scope of which is indicated by his list of research papers. Stephen is especially noted for his development of the BM25 ranking model, which built on the work of Karen Spärck-Jones on the term frequency.inverse document frequency model.

If you want to choose a date to mark the beginning of commercial enterprise search then 1970 is that date. It marked the launch by IBM of STAIRS (Storage and Information Retrieval System), an evolution of the AQUARIUS software that IBM developed to cope with the documentation for the defence of an anti-trust suit in the USA that started in 1969. STAIRS was specifically designed for multi-user time-share applications (the typical enterprise scenario) and remained on the IBM product list until the early 1990s. Jumping out of any sort of chronology in 1985 STAIRS was subject to a very thorough evaluation which raised doubts about the effectiveness of full text indexing. A review article by David Blair (Blair 1996), is a must-read for anyone with an interest in enterprise search and evaluation as it looks back at the 1985 evaluation with the benefit of substantial hindsight, and benefits from the fact that although Blair was one of the authors of the original review it comes across as an independent and unbiased assessment.

In the Conclusions section, Blair states:

"We have shown that the system did not work well in the environment in which it was tested and that there are theoretical reasons why full-text retrieval systems applied to large databases are unlikely to perform well in any retrieval environment."

By the mid-1970s mini-computers were being adopted very widely, and many organisations and companies saw this as an opportunity to develop text/document retrieval software products for these mini-computers. These included BASIS (Battelle Institute) and INQUIRE (Infodata).

So far this history has been dominated by developments in the USA but the mini-computer market stimulated software development in the UK, including ASSASSIN (ICI), STATUS (Atomic Weapons Research Establishment), CAIRS (Leatherhead Food Research Association) and DECO (Unilever). (I had a role on the development team of DECO from 1979-1981 which gave me a very valuable insight into the programming of search applications.) These and other applications all emerged towards the end of the 1970s. An interesting comparative review of them by John Ashford (a highly respected consultant) was published in 1984 (ashford 1984). These applications all evolved from specific organisational requirements which were then productised for use more widely, demonstrating that you did not need to be a large academic institution or software company to develop retrieval software. These systems were accessed through networked terminals; the IBM PC was not launched until 1981. The scale of the development of these products can best be assessed from A Technical Index of Interactive Information Systems, published as Technical Note 819 from the National Bureau of Standards in 1974. This report provides brief details of almost 50 software products.

The first Association for Computing Machinery (ACM) conference on information retrieval took place in 1971. The 1st Annual International SIGIR Conference on Information Storage and Retrieval took place in 1978. In 1979 the Institute of Information Scientists organised a two-day conference held at the Royal Society, London, entitled Computer Packages for Information Storage and Retrieval. The event attracted over 200 delegates.

As a footnote to this section on the 1970s it is important to highlight that the first assessment of the potential role of artificial intelligence in information retrieval was published in 1976 (Smith 1976). Just over a decade later Verity, the prototype for all enterprise search applications, emerged from a company specialising in AI development.

References

Ashford, J. (1984). Information storage and retrieval systems on mainframes and minicomputers: a comparison of text retrieval packages available in the UK. *Program: electronic library and information systems*, 18(2).

Blair, D. (1996). STAIRS Redux: thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science* 47(1), 4-22. https://yunus.hacettepe.edu.tr/~tonta/courses/spring2008/bby703/Blair.pdf

Manning, C.D., Raghaven, P. & Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Smith, L. (1976). Artificial intelligence in information retrieval systems. *Information Processing and Management*, 12(3), 189-222.

Spärck Jones, K. (2006). Information retrieval and digital libraries: lessons of research. Proceedings of the International Workshop on Research Issues in Digital Libraries (IWRIDL 2006).

5. 1980 – 1989 Rapid evolution

In the early 1980s there was a great deal of interest in the UK around the use of text retrieval software running on mini-computers. The STATUS User Group in particular was very active. These vendors were not especially interested in the commercial success of their products as the development had been justified on the need to meet internal information searching requirements within the organisation.

In the UK the Institute of Information Scientists played a very important role in stimulating interest in the capabilities of these applications through a series of Text Retrieval conferences between 1980 and 1990. The proceedings of these conferences make fascinating reading though sadly none are available in a digital format and only the 1998 and 1999 conference proceedings were published. However, most of the conference documents are held by the British Library.

As far as the technical development of enterprise search was concerned probably the most important advance was the release of the Snowball English language stemmer developed by Dr. Martin Porter. To be pedantic it was first released in 1979 but was not widely promoted until 1980. Martin Porter tells the story from a 2001 perspective on his website where his original stemming code and many more algorithms for various languages are available as open source. According to the Wikipedia entry the name Snowball was chosen as a tribute to the SNOBOL programming language, with which it shares the concept of string patterns delivering signals that are used to control the flow of the program.

Martin Porter, together with John Snyder, also developed the Muscat (MUSeum CATalogue) search application while at Cambridge University. Released in 1984 it sought to take advantage of the work of Stephen Robertson and others on a probabilistic approach to information retrieval. Muscat Ltd. became a successful company with clients that included Fujitsu, the Japanese IT company. Muscat was eventually rewritten and released as the open source Xapian library which survived the eventual acquisition of Muscat Ltd. by a short-lived dotcom era company and is still available. There is a good summary of Muscat on the Flax website.

By the mid-1980s the IBM STAIRS full-text search application was setting the standard for enterprise search. In 1985 a wide-ranging research study was carried out by Blair and Maron of the retrieval performance of STAIRS, which at that time was being promoted as a litigation support tool. The results were far from impressive (Blair and Maron 1985). This study remains the most comprehensive of its type, with nothing approaching it having been published in the last thirty years. It had commercial implications for the legal sector as this was the time when there started to be a number of major anti-trust cases brought by the US Department of Justice where reliable access to millions of corporate documents was of great importance. It should also be borne in mind that the IBM PC had been launched in 1981 and it was during the 1980s that documents started to be created on personal computers rather than being transcribed onto word processors.

I would suggest that the first commercial enterprise search application other than STAIRS was developed by Fulcrum Technologies, established in Ottawa in 1983. This was a client-server application, rather than mainframe and offered the first API for writing information retrieval applications. It was most visible for the rest of the decade as a provider of search software for CD-ROM applications. From 1983 to 1988 Fulcrum pretty much had the search market to itself but failed to make much headway. The arrival of Verity (see below) born in the entrepreneurial climate of California, marked a gradual decline of Fulcrum as a business. A succession of owners over the 1990s led eventually to Fulcrum being purchased by Hummingbird in 1997, which itself was then acquired by OpenText in 2006.

In 1985 Advanced Decision Systems was set up in San Jose, California with the objective of developing expert system and artificial intelligence applications. In 1986 David Glazer and Philip Nelson developed an innovative search application called Topic which was beta tested with success by the US Strategic Air Command. Topic made use of a probabilistic search ranking engine which offered significantly better management of ranking than the Boolean operators that had been used prior to the release of Topic, though STAIRS also used this model. This early success led to the spin-out of what was to become Verity from ADS, led by Michael Pliner with a technical team led by David Glazer and Philip Nelson. There can be no doubt that Verity was the proto-typical enterprise search application as unlike IBM STAIRS it was platform agnostic. At launch a multi-user licence cost $39,500, quite a substantial licence fee in the late 1990s.

Two other search software companies started out towards the end of the 1980s. David Thede set up dtSearch in 1988, initially offering a desktop search application. dtSearch remains one of the very few search software vendors to have been in the same ownership from start-up to the present day. Also in 1988 but across the other side of the world in Australia Ian Davies was developing the Isys software suite. This ended up being acquired by Lexmark in 2012. Several others were on the drawing board but did not emerge until the early 1990s.

The decade also marked the birth of a project at CERN in Switzerland to create what would become the World Wide Web. Tim Berners-Lee submitted his report Information Management – A Proposal in March 1989. It is important to appreciate that the initial purpose of the project was to be able to search through CERN documentation and thereby an enterprise search project was the start of the global web search business. W3C has compiled a very useful chronology of the subsequent development of the World Wide Web over the period from 1989 to 1995.

Reference

Blair, D.C. and Maron, M.E. (1985) Communications of the ACM 28(3), 298-299. https://doi.org/10.1145/3166.3197

6. 1990 – 1999 Innovation in retrieval technology

Before looking at the enterprise search business itself there were important developments in the understanding of how people searched, and in novel technical advances in search. Marcia Bates started to make us think about search behaviour in her 1989 paper on berry picking as a metaphor for the process of discovery. Peter Pirolli's work on information foraging was published in 1999. Although this is right at the very end of the decade being covered it is indicative of the research that was being undertaken looking at information systems from a user behaviour perspective, with Jakob Nielsen (the founder with Don Norman of the Nielsen Norman Group waiting in the wings at Sun Microsystems from 1994 to 1998. From an enterprise search perspective the work that was undertaken at the University of Huddersfield by Stephen Pollitt on faceted navigation was ground-breaking. The concept was taken up and developed further by Marti Hearst with her Flamenco project.

From a technical perspective the challenges of indexing and searching the World Wide Web were now starting to be addressed, taking search in some very different directions. Alta Vista was not the first WWW search engine but the team working on it gained an immense amount of knowledge about web crawling and indexing at scale. Two members of the team founded Exalead in 2000. Google followed in 1998 and of course the arrival of enterprise web applications such as intranets opened up a potentially very large market for enterprise-level search. Sadly the IBM HITS algorithm (later integrated into the IBM Clever project) didn't have a chance against the Google PR machine. During the late 1980s and then into the 1990s advances in natural language processing were rapid as machine learning approaches and developments in machine translation opened up new opportunities for search. Latent

Semantic Analysis first emerged in 1988 and Probabilistic Latent Semantic Analysis in 1999, the latter forming the basis of the Recommind e-Discovery application, now owned by OpenText. Lucene, written by Doug Cutting, also appeared in 1999. This was (and remains) a free open-source search engine software library and is now widely used in conjunction with Solr (developed by Yonik Seeley), ElasticSearch and Lucidworks, amongst many others.

The stage was set for the emergence of a significant number of search vendors. Verity was gaining momentum but finding it difficult to achieve profitability. In 1993 RetrievalWare emerged and started a trend for search software companies to have multiple owners. How it ended up in FAST Search and Transfer via Excalibur is, to say the least, complicated.

The Infoseek/Ultraseek/Inktomi/Verity/Autonomy saga, which started in 1993, was yet another complicated journey. Interestingly Ultraseek was branded as Ultraseek Enterprise Search and by the time it was acquired by Autonomy had around 15,000 customers. Verity achieved an IPO in 1995, achieving funding of $40m, double the amount anticipated. This probably encouraged (at least indirectly) the arrival of Autonomy (1996), FAST Search and Transfer (1997) and Endeca (1999).

The development of the enterprise search business in the early 1990s is not well documented. Many of the entrepreneurs who had a vision for search have been interviewed by Stephen Arnold in his invaluable Wizards Index column. In the paragraph above most of the links are to Wikipedia entries, which inevitably vary in quality and depth but hopefully are at least a starting point for research. The distinguished journalist and philanthropist Esther Dyson tracked the development of internet companies during this period.

7. 2000 – 2009 The start of industry consolidation

The decade from 2000-2009 was marked by the high visibility of Verity, Autonomy and FAST Search and Transfer and the beginning of consolidation in the search business. Verity grew rapidly over the period from 2000-2005 and started to achieve a respectable level of profitability. Revenues in 2003 were just over $100 million. These increased to $150 million by 2005 with the company sitting on around $250m in cash and investments. Autonomy acquired Inktomi (or rather the Ultraseek product) in 2003 and Cardiff Software in 2005. By late 2005 there were 160 employees and the company claimed that 15,000 companies and other organisations had licensed its software.

A potential game-changer emerged in 2002. This was the Google Appliance, which was a substantial amount of Google technology delivered on a Dell server in a yellow casing designed to be inserted into a standard server rack. The pricing model was document based, but this came with some hidden implications, notably calculating the cost of Excel files based on the number of worksheets. For CIOs that had long argued for an enterprise search that worked like Google it was an answer to their dreams. Google increased the size of the server configuration and released a number of software upgrades. At first the reaction was very positive but it was not easy to optimise the search results and the level of support from Google was very limited.

Over the same period of 2003-2005 FAST Search and Transfer revenues increased from $42 million to over $100 million, but the company had over 450 employees and the 2005 Annual Report is a tale of woe about a whole range of investments and other transactions. The FAST IPO had taken place in 2001. The company then sold off its web search interests in 2003, including AllTheWeb which has now reappeared as a component of Vespa. The acquisition of RetrievalWare followed in 2007 but there were already concerns about the way in which the company was presenting its accounts.

In 2000 Autonomy raised $124 million of investment funds when it floated on NASDAQ and then in 2003 started the process of acquiring a substantial stable of companies, starting with the video software company Virage in 2003. Then in 2005 it acquired Verity for $500 million, a significant multiplier on $7 million net income. By 2006 Autonomy was reporting revenues of $250 million but probably half of this amount was generated by Verity. Over the next three years Autonomy also acquired Blinx, Zantaz, Merido and probably most remarkably Interwoven, a WCMS vendor. In 2008 Autonomy became a member of the FTSE100, and by 2009 the company had revenues of £740 million and over 1600 employees.

The acquisition of FAST Search and Transfer by Microsoft in 2008 came as a surprise, as did the purchase price of $1.2 billion. It seemed to suggest that Microsoft was going to be an enterprise search provider, based around the very powerful FAST ESP search platform. However within months of the acquisition concerns were being raised about the extent to which the booked revenues of FAST Search and Transfer were being recognised, a situation that also arose in 2011 with the HP acquisition of Autonomy. One day the full story of both acquisitions may emerge. In the event Microsoft stripped out elements of FAST ESP and incorporated them into the FAST Search Server for SharePoint 2010. Such was the reputation of FAST that many organisations were under the impression that they had actually acquired the ESP product bundled into SharePoint.

Although Verity, FAST and Autonomy were the most visible enterprise search applications others were also being developed quite successfully, including Endeca, Exalead, Vivisimo, ISYS Search and a number of others, but their independent existence continued for a few more years. Of particular note was P@noptic which developed from a research project dating back to 1991 at the Commonwealth Scientific and Industrial Research Organisation, the national research organisation in Australia. When Google arrived on the search scene CSIRO saw an opportunity to address the problems of enterprise search by commercialising this ongoing research into text retrieval and from it created P@noptic. This very capable search application had a number of very neat technical elements and quite quickly gained a collection of highly satisfied users from operations in Australia, the USA, the UK and Poland. The company was spun off in 2005 as Funnelback Pty Ltd and was sold to Squiz in 2009. A history of the project has been published as an autobiography by David Hawking, the leader of the CSIRO project team and who was then actively involved in the commercialisation of the technology in Funnelback.

8. 2010 – 2019 Rebranding enterprise search - 'cognitive search' and 'insight engines'

From 2010 to 2013 there was a rapid consolidation in the enterprise search business. Between 2010 and 2012 Exalead was acquired by Dassault (2010), Autonomy by Hewlett Packard (2011), Endeca by Oracle (2011), Vivisimo by IBM (2012) and ISYS Search by Lexmark (2012). Some of these vanished without trace, some notionally exist (Exalead) and Autonomy returned to the UK following its acquisition by Micro Focus. In August 2022 Micro Focus was acquired by Open Text, a Canadian company with diverse enterprise applications including enterprise search.

Others emerged to fill the gaps. As mentioned above Funnelback was initially developed by CSIRO in Australia but did not really move into the limelight until the establishment of a UK office in 2009 following its acquisition by Squiz. Lucid Imagination was set up in 2009 and was then renamed LucidWorks in 2012. BAInsight dates back to 2003 as a supplier of add-on modules to SharePoint but over the last few years has repositioned itself as more of a systems integration company and in 2021 was acquired by Upland Software. Mindbreeze, an Austrian company offering a search appliance, was founded in 2005 and as with the other companies mentioned above has flourished over the last few years.

In 2016 Google announced it was leaving the enterprise market and terminated the licenses at the end of 2018 without offering a replacement product.

Looking back at the Gartner Magic Quadrant for Information Access Technology in 2005 there were four companies in the Leader/Visionary Quadrant, and they were FAST, Autonomy, Verity and Endeca. The majority of the companies surveyed in 2005 were towards the lower end of the Ability to Execute axis, and that has always been a challenge for the enterprise search business. Many companies with very good technology could not generate sales and cash flow to finance the marketing and sales effort needed to get to a critical mass. Over the last decade the market has been dominated by Microsoft SharePoint in terms of an installed base of search functionality (perhaps close to 300,000 installations?) though Google built up a substantial installed base of appliance servers before leaving the stage. The Enterprise Search Summit was launched in New York in 2008 and the exhibition space was full with around 40 vendors. Those were the days!

The Enterprise Search Europe event was launched in 2011 but 2015 marked its closure as there were just not enough sponsors to keep the delegate fee at a sensible level. Thanks to Findwise we do now know much more about the way in which enterprise search is being implemented and used through the Enterprise Findability Surveys that started in 2011 and continued to 2016. The survey was run again in 2019 and the Danish consulting company IntraFind introduced its digital benchmarking service. Both surveys confirm that there have been no improvements in the low levels of search satisfaction.

Over the last few years the concepts of ‘cognitive search’ and ‘insight engines’ have been proposed by two IT industry analysis firms, Gartner and Forrester. The basis of both is that search results can be customised down to the level of an individual employee based on what they are working on within the context of their colleagues. The aim of these applications is to deliver the most relevant information at position 1 on a search results page with the searcher just entering anything from a single word to a section of text they are working on. The technology involved is a combination of AI and machine learning allied to developments in natural language processing.

As yet there is no independent research that shows whether these approaches are scalable and extensible for enterprise-wide use in situations where employees are working on multiple tasks and projects simultaneously and in a range of languages.

9. 2022 Defining 'good practice' in enterprise search

The first published bibliography of research into information retrieval was published in 1964 (Snoddy 1964) and covered the period from 1957–1961. Fast forward to 2021 when the IR Anthology was established with a database of over 40,000 papers (Potthast 2021). However, this collection is not fully comprehensive in its scope, and in total it could be that there are approaching 100,000 research papers.

In the case of academic research into enterprise search there has been only one research paper published which considers the way in which enterprise search is used across a single organisation. This paper (Lykke 2021) provides a wealth of data on how employees make use of enterprise search. The organisation was a Danish biotech company with 7500 employees. With any individual case study the issue is always the extent to which the outcomes scale to other organisations. Without going into analytic detail it is reasonable to assume that it does scale, certainly to other medium to large-scale high-technology organisations.

There have been a number of research papers which document the outcomes of projects to assess the way in which specific groups of employees (such as engineers) make use of enterprise search applications in a number of different organisations. Cleverley and Burnett at Robert Gordon University, Aberdeen, have published a number of papers on the way in which enterprise search has been used in a specific large oil and gas company. In particular in 2018 they identified that the root causes of search dissatisfaction were problems with the technology implementation, the quality of the content and the extent to which users were trained.

In 2019 the two authors published an excellent overview of enterprise search based on interviews with vendors and users.

Two other related areas where there has been a reasonable amount of research is the use of internal search applications by professional users, such as lawyers, patent agents, recruitment agents and clinical staff, and analyses of the way in which search is a core component in the successful completion of a task. In the case of professional users the research indicates that there are some significant differences in the use of specific features of the user interface.

Taking into account the research papers cited in the 2018 and 2021 studies it would indicate that there are probably fewer than 50 peer-reviewed papers into enterprise search-related topics despite the fact that millions of employees around the world use these systems to undertake business-critical searches. Each week there are around 200 research papers published in the IR section of arXiv but only a few have any relevance to the document-centric repositories that dominate enterprise information resources.

There are a number of reasons for the lack of interest by academic research teams in enterprise search. These include:

- The concern of organisations that the research will reveal its strengths and weaknesses
- During the duration of a typical three-year PhD study there could be very significant changes in business direction that might invalidate the research, for example an acquisition, divestment, or the establishment of a new line of business
- Because enterprise search is security trimmed to ensure that only employees with appropriate access permissions see certain information it is very difficult to know whether the inability of an employee to find information is actually an outcome of a security barrier
- The dominant ‘enterprise search’ application is Microsoft Search but this is an atypical example as it is specifically designed to work within the Microsoft technical architecture. As a result, *inter alia* snippets are not well-presented and the search analytics applications are very limited.

- There are currently no undergraduate or graduate courses in enterprise search technology and management in North America or Europe. This means that there is very limited knowledge of enterprise search within even the Information School community and no incentive to undertake search to enhance the visibility of a department. As far as the IT management community is concerned it is of note that neither the British Computer Society nor the Association for Computer Machinery in the USA has published a book on enterprise search.
- The standard academic career progression of PhD, post-doctoral research, lecturer and upwards to potentially Professor does not accommodate time spent in a corporate enterprise search role which would not count towards academic advancement.

Another issue that faces enterprise search managers is that there are no enterprise search conferences at which good practice can be formulated and shared. At one time there was an Enterprise Search Summit in the USA but this became just one of many tracks at the annual KM World conference held in Washington D.C. each year. An Enterprise Search Europe event was launched in 2011 but was discontinued in 2014. The primary reason was a lack of sponsorship support from vendors who felt that this was not a good use of the time of their sales teams.

The Information Retrieval Specialist Group (IRSG) of the British Computer Society does include enterprise search in the scope of its annual one-day Search Solutions Conference and there is an Industry Day at the European Conference on Information Retrieval (also managed by IRSG) but as with the Search Solutions Conference there is no specific focus on enterprise search.

References

Snodey, S.R. (1964). Information retrieval – a comprehensive indexed bibliography of 1957-1961 world literature. *IEEE Transactions on Engineering Writing and Speech*, 7(1), 22-38.

Potthast, M. et al. (2021). The information retrieval anthology. SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. July. 2550-2555. https://dl.acm.org/doi/10.1145/3404835.3462798

Lykke, M. et al. (2021). The role of historical and contextual knowledge in enterprise search. *Journal of Documentation*. https://kbdk-aub.primo.exlibrisgroup.com/discovery/openurl?institution=45KBDK_AUB&vid=45KBDK_AUB:AUB&sid=pureportal&doi=10.1108%2FJD-08-2021-0170

Cleverley, P. & Burnett, S. (2019). Enterprise search and discovery capability: the factors and generative mechanisms for user satisfaction. *Journal of Information Science*, 45(1), 29-52. https://doi.org/10.1177/0165551518770969

Cleverley, P. & Burnett, S. (2019). Enterprise search: a state of the art. *Business Information Review*, 36(2), 60-69. https://doi.org/10.1177/0266382119851880

10. Lessons learned

In 2012 I undertook a research project for the European Commission to assess the opportunities and barriers that enterprise search software companies faced in building successful businesses. The project was stimulated by the acquisitions that had taken place over the previous few years and the concern of the Commission that the EU might be dependent on US companies for enterprise search software.

The barriers have remained largely unchanged. The most important barrier is that there is a significant shortage of people with the skills to support the development, installation and management of enterprise search applications. Experienced search managers can command high salaries and their organisations will make every effort to retain them as finding someone to replace them, especially with similar expertise and experience, is going to be very difficult.

The second barrier is that product differentiation is very difficult to achieve. At the present time search software vendors are promising AI/ML magic with no evidence as to the eventual performance. It is very difficult for them to build a business case for investment in their software, and also difficult to invest ahead of demand in the systems integration skills needed to implement enterprise search, especially in multi-national organisations.

The major change has been the market dominance of 'enterprise search' by Microsoft. The Microsoft applications are optimised to provide effective search of Microsoft files in Microsoft repositories, with the exception of Azure Cognitive Search. In effect the Microsoft search functionality is free because it is an element of Microsoft 365 which dominates the office productivity market. Replacing this search with either another commercial or an open source product is very difficult as an element of the business case has to be a justification of why investment needs to be made in replacing the ‘free’ Microsoft search application. SharePoint search may well be seen as ‘good enough’ by a CIO as they have no experience of other search applications and improving search, with its very limited immediate impact, comes way down the priority list for investment.

The only public search companies are ElasticSearch, which is open source, and Coveo, which now specialises in e-commerce search. Indeed e-commerce search is very much in demand because it is easy to make a business case around increased sales and customer retention.

In total there are over 70 companies offering enterprise search software. Most are small businesses that focus on their national market, especially in the USA. They are all funded by venture capital, and investors are always looking for a return on their investment. The only exit strategy investors have available are to sell the technology to a larger company, which is what happened with the companies acquired in the late 2010s. There was no value in the client base. The most visible effect of this technology acquisition is the case of Attivio, which sold its technology IP to ServiceNow with the result that Attivio was not able to continue in business and many clients were left without an immediate replacement option.

11. The end of transparency?

The major technology advance over the period from the initial availability of enterprise search applications in the 1980s running on mini-computers until around 2020 was arguably the gradual introduction of the BM25 ranking model from around 2010 to replace TF.IDF. There have been many variants of BM25 but it became the default ranking model for most enterprise search applications.

Search can fail in many ways, as outlined in a schematic from Clearbox Consulting. From a user perspective it is often difficult to understand why a search has returned a poor set of results with low relevance to the query, if indeed it returns any results at all. Self-diagnosis is impossible, which is one of the reasons that successful enterprise search applications invariably have a strong search support team that is proactive in ensuring that search is satisfactory.

Surveys over the last decade have all indicated that perhaps only 20% of organisations have a search application that delivers a high level of search satisfaction. In the course of writing this book the author took part in the Intranet Italia Day conference in Milan in May 2022. When the audience of over 150 intranet managers was asked to raise their hands if they knew that employees were satisfied with the search performance of their intranet only five delegates did so.

The use of BM25 and related models for ranking does make it possible to reverse engineer a query and results to understand what the possible causes of the poor performance might be. Search applications have dashboards that can then be used to boost particular words or phrases, and it is also possible to manually ensure that entity extraction and name similarity routines are working effectively.

With the arrival of machine learning, dense vectors, neural networks and very large pre-trained language models the transparency of the search process disappears. A core requirement of enterprise search is that employees trust it because search failure in any degree could put the organisation, and their own careers, at risk through making a flawed decision on the basis of not finding relevant enterprise-created information.

The aim of AI-based search is the Holy Grail of understanding the intent of the query in order to deliver the most relevant set of results. No research has ever been undertaken to categorise enterprise intents. Research into the intents behind web search queries suggests that the range of intents, and the difficulty of categorising them, are quite considerable.

At present AI-based search is in the hype-stage of development, which experience shows is then followed by a period of disillusion with the initial promise of the technology. Out of this disillusion comes a reality check and a gradual period of wide-spread adoption with significant benefits to the organisation and to the individual employee. Even with the benefit of 60 years of search development it is not possible to put a time-scale on this evolution or to forecast when it might be of value to write a second edition of this history.

Appendix - Research Sources

The evolution of enterprise search is quite complicated and poorly documented. In this report I have set out a few of the milestone events and developments. They are a personal selection of history highlights and I make no attempt at being 'comprehensive'. It is 'A history' and not '*The* history'.

The functionality that is now encapsulated in enterprise search software has been in constant development since the early 1950s, with especially rapid evolution in the 1970s and 1980s with the availability of large-scale commercial online search services such as Lockheed Dialog, SDC Orbit, BRS and Mead Data Central (Lexis). I started my career in search in 1976 and have had the good fortune to have met many of the early pioneers, notably Roger Summit, Charles Bourne, Carlos Cuadra, Jerry Rubin, Noel Isotta, David Raitt and Cyril Cleverdon. Whilst working in Cupertino in the early 1980s I also had the opportunity to meet research staff from Stanford Research Institute who had worked with Doug Engelbart. Other personal milestones include working on the development of one of the early UK enterprise software applications (DECO) in 1981-1982 and in 1983-1984 inadvertently playing a role in the establishment of OpenText a decade later when I was involved with the conversion of the Oxford English Dictionary into a machine-readable format for editing and production.

Any history of enterprise search is intrinsically linked to the history of information retrieval, a term first used by Calvin Mooers in 1950. There have been many articles published on the history of information retrieval but by far the most readable is the chronology of information retrieval research written by Mark Sanderson and W Bruce Croft. I've always been intrigued that the URL id is 1066 and have often wondered if that was an accident or by design!

In 2019 Donna Harman published Information Retrieval: The Early Years, combining a very comprehensive bibliography of almost 300 research papers with her own experience of having been at the forefront of IR research. However, there are no specific references to enterprise search. The role of the Chemical Abstracts Service in advancing the use of computers in information retrieval through the commitment of James Perry, G. Malcolm Dyson and Pete Luhn is not mentioned at all.

I authored a profile of G. Malcolm Dyson for the RSC CICAG Newsletter published in late 2021. This focused on his work in the area of chemical information management. A more detailed biographical account of his life is in preparation.

Anyone writing a history of enterprise search is enormously indebted to Charles Bourne and Trudi Bellardo Hahn for their book A History of Online Information Services 1963-1976. Their book also provides a substantial amount of detail about enterprise search applications, though this term was not used at the time.

Another excellent source is a literature review entitled Cooperation, Convertibility, and Compatibility Among Information Systems: A Literature Review published in 1966 by the US Department of Commerce that considered the issues arising from a multiplicity of information systems even at that early stage of development. This review provides a very good outline of the development of computer-based information services dating back to the early 1950s as well as reflections on scientific communication in the widest sense from the founding of the Royal Society in London in 1660.

Stephen Robertson contributed a survey on Computer Retrieval as seen Through the Pages of Journal of Documentation to B.C. Vickery, Ed., Fifty years of information progress: a Journal of Documentation review. London: Aslib (1994) . It contains a bibliography of 146 items. Brian Vickery's career spanned much of the period covered in this history and his personal account of his work provides valuable insights into events in both the USA and the UK with regards to the role of computers in information retrieval. Stephen Robertson has also published B C, Before Computers: On

Information Technology from Writing to the Age of Digital Data in 2020.

Another personal perspective on the development of search technology has been written by David Hawking, who took a lead role in the development of the P@ntopic search application which was later commercialised as Funnelback.

Probably the definitive text on information seeking in its broadest sense is Looking for Information by Donald Case and Lisa Given, which has a very good section on the models that have been developed to help define and manage the process of information search.

Jeremy Norman's History of Information web encyclopedia, can be browsed through chronologically. A series of interviews with the pioneers of the pre-internet online search services was published in the Searcher magazine and these are an invaluable source of primary information on these services.

Information – A Historical Companion was published in 2021, edited by Ann Blair, Paul Duguid, Anja-Silvia Goeing, and Anthony Grafton. This 880-page book is a comprehensive account of the development of information handling from the early dynasties of China onwards. Chapter 13 is specifically about search and includes a useful bibliography.

Also of immense historic value are the series of interviews carried out by Stephen Arnold between 2008 and 2013 and published in his Wizards Index. Most of the founders of enterprise search applications tell the inside stories of how they created, launched and developed these applications.

All the links were checked on 20 June 2022. Please report any broken links to oer@sheffield.ac.uk.

I would of course appreciate comments on factual inaccuracies, omissions and additional resources. Contact me.