



Search Insights 2019

The Search Network

March 2019



Contents

● Introduction	1
● 2018 in review	5
● The rise of the relevance engineer	8
● Rethinking 'advanced search': a new approach to complex query formulation	12
● Search insights - a view from a search manager	16
● Microsoft search	19
● Why we (still) need taxonomies (and the taxonomists who nurture them)	23
● The impact of corporate language policies on enterprise search	26
● Enterprise search failure	29
● Content audit	33
● Search as a service	38
● Search team skills	41
● Achieving enterprise search satisfaction	44
● Appendix A Enterprise search software	49
● Search resources: books and blogs	51
● Glossary	53

This work is licensed under the Creative Commons Attribution 2.0 UK: England & Wales License. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc/2.0/uk/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Editorial services provided by Val Skelton (val.skelton@blythespark.co.uk)
Design & Production by Simon Flegg - Hub Graphics Ltd (www.hubgraphics.co.uk)

Introduction

Martin White

The Search Network is a community of expertise. It was set up in October 2017 by a group of eight search implementation specialists working in Europe and North America. We have known each other for at least a decade and share a passion for search that delivers business value. We are increasingly concerned by a focus on search technology by vendors that takes no account of business requirements, implementation challenges and the need for a skilled support team. Search is not a product or a project. It requires an on-going commitment to support changing user and business requirements and to take advantage of enhancements in technology.

Members of the Network have web site search, enterprise search and search application development expertise with on-premise, hybrid and cloud implementations. We all work as individuals or micro-companies and have no commercial relationships with any search vendor or implementation partner. We often assist in identifying vendors for evaluation and consideration.

Some of us have experience with commercial vendors (including SharePoint) and others work exclusively in the open source search business. We recognise that the best option is the one that most closely meets the requirements of the organisation. Often these requirements involve members of The Search Network bringing in colleagues with specific skills or to extend our geographic scope. The Search Network is an informal community, not a hub-and-spoke network. You can talk to any one of the members and they can bring in others as appropriate. We look forward to helping you achieve search excellence.

In the course of our work we have gained a substantial amount of experience which can be matched by very few IT managers. In total the contributors to Search Insights 2019 have well over 50 years of experience in helping organisations to find business-critical information, working with enterprise search, e-commerce and web site search, and with specialised search applications. Not only do we work with different types of search applications, but we also write in our own style and from our own individual experience. Our objective in writing this report is to summarise some of the insights we have gained from these projects and make this knowledge open to the search community worldwide. That is why there is no charge for this report, and it carries no sponsorship.

Our most significant contribution to our clients is a very good understanding of what an effective search application can deliver in terms of business benefits and employee engagement. Very few organisations have had an opportunity to see and use the range of search applications that we have worked on.

We hope that you will find that Search Insights 2019 (together with Search Insights 2018 [Search Insights 2018, published in April 2018](#)) will enable you to make the right decisions about providing your organisation with effective access to information.

David Hobbs, [David Hobbs Consulting](#) (USA)

David helps organisations make higher impact digital changes, especially through the early development of a strategy to best frame these initiatives before they begin. He is the author of Website Migration Handbook and Website Product Management. His clients include the Center for Internet Security, the Library of Congress, the Mideast Broadcasting Company and the World Bank. Follow David on Twitter [@jdavidhobbs](#).

Charlie Hull, [Flax](#) (UK)

Charlie is the co-founder of Flax, which builds open source search and Big Data solutions for clients worldwide. He writes and blogs about search topics, runs the London Lucene/Solr Meetup and regularly speaks at, and keynotes, other search events across the world. He co-authored Searching the Enterprise with Professor Udo Kruschwitz. Follow Charlie on Twitter [@o19s](#).

Miles Kehoe, [New Idea Engineering](#) (USA)

Miles is founder and president of New Idea Engineering (NIE) which helps organisations evaluate, select, implement, and manage enterprise search technologies. NIE works and partners with most major commercial and open source enterprise search and related technologies. He blogs at [Enterprise Search Blog](#) and tweets as [@miles_kehoe](#), [@Ask Dr Search](#) and [@SearchDev](#).

Helen Lippell (UK)

Helen is a taxonomy consultant. She works on taxonomy development projects, including taxonomy audits, ontology modelling, tagging initiatives, semantic publishing, metadata training and more. Her clients include the BBC, gov.uk, Financial Times, Time Out, RIBA and the Metropolitan Police. She writes and speaks regularly and is the programme chair of Taxonomy Boot Camp London. Follow Helen on Twitter [@octodude](#).

Agnes Molnar, [Search Explained](#) (Hungary)

Agnes is the managing consultant and CEO of Search Explained. She specialises in information architecture and enterprise search. She shares her expertise on the [Search Explained](#) blog and has written and co-authored several books on SharePoint and Enterprise Search. She speaks at conferences and other professional events around the world. Follow Agnes on Twitter [@molnaragnes](#).

Eric Pugh, [OpenSource Connections](#) (USA)

Eric is co-founder and CEO of OpenSource Connections where he helps federal, state and commercial organisations develop strategies for embracing open source software. He co-authored Enterprise Solr Search, now in its third edition. He is interested in how Search is being invigorated by Machine Learning and exploring approaches for sharing data the way the open source movement shares code. You can follow him on Twitter at [@dep4b](#)

Doug Turnbull, [OpenSource Connections](#) (USA)

Doug is CTO of OpenSource Connections and the author of Relevant Search. His goal is to empower the world's best search teams. He has assisted with search at organisations in a variety of domains. His clients include Wikipedia, Snagajob, Careerbuilder, and many search organisations. Follow Doug on Twitter [@softwaredoug](#).

Martin White, [Intranet Focus Ltd](#) (UK)

Martin is an information scientist and the author of *Making Search Work* and [Enterprise Search](#). He has been involved with optimising search applications since the mid-1970s and has worked on search projects in both Europe and North America. Since 2002 he has been a Visiting Professor at the Information School, University of Sheffield and is currently working on developing new approaches to search evaluation. Follow Martin on Twitter [@IntranetFocus](#).

Guest contributors

Jon Chamberlain, [University of Essex](#) (UK)

Jon is a lecturer in Human-Computer Interaction with experience of industrial and academic computer applications (language processing, game design, social network analysis) in the domains of citizen science, marine conservation and human rights observation. He was the lead developer of the Phrase Detectives crowdsourcing project that has created the largest collaboratively-produced resource for anaphoric reference and continues to investigate using crowdsourcing through games in the 5-year DALI project. His recent research looks at using crowdsourcing to collect labels for images collected from underwater drones for photogrammetry and navigation.

Sam Marshall, [ClearBox Consulting](#) (UK)

Sam is the owner of ClearBox Consulting and has specialised in intranets and the digital workplace for over 19 years, working with companies such as GSK, Vodafone, TUI Travel, Sony and Unilever. His current activities focus on intranet and digital workplace strategy, and the business side of Office 365. He is the lead author of the annual 'Share-Point Intranets in-a-box report', the most comprehensive review of these add-on tools. Sam is a regular keynote speaker at international conferences and has been named a 'Contributor of the year' for his CMSWire column four years in a row. In 2015 Sam was given the Intranet Now award for 'remarkable contribution to the intranet community'. Follow Sam on Twitter [@sammarshall](#).

Karen Renshaw, [Grainger Global Online](#) (UK)

Karen is passionate about improving the onsite search experience, to demonstrate that investing in the experience can benefit the overall customer journey and core business metrics. She has over ten years' experience of managing search teams to drive business benefits. She is currently Head of Search and Content for Grainger Global Online. Previously, Karen was Head of Search for RS Components where she set up a Global Search team and developed a search migration programme.

Stephanie Segura Rodas, (Lima, Peru)

Stephanie has worked in information security roles in a number of international companies. She has a degree in systems engineering from the University of San Martin De Porres, Lima, Peru and was awarded a Distinction for her MBA from the Management School, University of Sheffield in 2018.

Tony Russell Rose, [UXLabs](#) (UK)

Tony is founder of 2dSearch, a start-up applying artificial intelligence, natural language processing and data visualisation to create the next generation of advanced search tools. He is also director of UXLabs, a research and design consultancy specialising in complex search and information access applications. Previously Tony has led R&D teams at Canon, Reuters, Oracle, HP Labs and BT Labs. He currently holds the position of RAE Visiting Professor of Cognitive Computing and AI at Essex University and publishes widely on information retrieval, NLP and human-computer interaction. He has a PhD in Computer Science and a first degree in engineering, majoring in human factors. Follow Tony on Twitter [@tonygrr](#).

2018 in review

Martin White

Google: back in the enterprise search business

In September 2018 Google announced a [cloud enterprise search](#) solution that will offer a range of connectors to facilitate a federated search across multiple applications. Technically, this is a very elegant one-stop approach, but the big unanswered question is what the pricing model is going to be. The Google enterprise search appliance offered a notionally transparent per-document pricing model, but many users found the definition of 'document' somewhat idiosyncratic.

Google's major problem will involve displacing existing applications. Proving that one search solution is better than another on a 'proof of concept' test is no guide as to how it will perform in a production environment. The inevitable change in the user experience alone will pose a significant challenge.

At present, pricing appears to be something agreed upon on an individual basis, probably linked to the publicity value of the customer. (I have lost count of the number of Workplace by Facebook presentations I have seen recently where no money has yet changed hands.) There are indications that it will be based on document volume and users, but very few organisations have any sense of the number of documents that they hold, and how many of these documents are anything more than interim versions that have little or no value.

Of course, the investment is not just in the software license but also in the integration support. Google lists almost fifty implementation partners but looking through their websites there seems to be little information about the extent to which they have enterprise search expertise. One of the challenges of enterprise search implementation is that the challenges at this scale are always multi-lingual and multi-country. Finding a single implementation partner with the optimum balance of skills and expertise is going to be difficult. Search implementation skills also need to be aligned to a good knowledge of the business sector and its processes.

Microsoft: unravelling a complicated search legacy

Microsoft acquired FAST Search and Transfer ten years ago and then buried it after stripping out some elements that ended up in FS4SP. In particular, Microsoft discontinued the content processing pipeline. At present, Microsoft has a complex range of search applications, including 'classic' and 'modern' search, Azure (which is Elastic under the covers) and Bing for Business. Complicating matters, Microsoft's announcements often lack clarity on whether it is a product or a promise.

Towards the end of 2018 there were announcements about a more coordinated Microsoft strategy that indicated there would be enterprise-wide solutions available during 2019. The focus is of course on searching Microsoft applications. How third-party applications are going to be implemented remains to be seen. I am curious to see how or if the underlying graph search technology will cope with indexing non-Microsoft applications. Unlike Google, Microsoft has to support a substantial legacy market and that limits its ability to throw bits of its technology stack away.

Elastic: strong IPO, but is it sustainable?

Elastic recently completed a very successful IPO, with the stock currently trading at almost double the offer price, valuing the company at around \$5 billion. The revenue forecast for April 2018 – March 2019 is around \$250 million and in October 2018 the company had cash and cash equivalents of \$318 million which could be used to fund further acquisitions. Total subscription customer count was over 6,300 and the total customer count with purchases of more than \$100,000 was over 340. Clearly the IPO has been very ‘successful’ but how long will it take for the investors to gather a return on the investment? The last three decades have been full of venture-capital funded search businesses, but I doubt any of them were profitable and the only winners have been instances where major IT companies have paid good money to gain access to technology, not to a client base. The acquisition by IBM of [Vivisimo](#) in 2012 is a good example.

Haystacks in the USA and Europe

In 2016 Doug Turnbull and John Berryman (Open Source Connections) wrote [Relevant Search](#) as a handbook for search managers and developers. Although written from an open source perspective the book is of value to anyone struggling to understand relevance management. Following up on the success of this book Open Source Connections launched the [Haystack Conference](#), which took place in Charlottesville, Virginia, in April 2018. The conference was very successful, with over 100 delegates.

The first [European Haystack Conference](#) took place in London in October 2018 and again attracted around 100 delegates. The next Haystacks Conference will take place in Charlottesville in April 2019, with a European event later in the year.

Relevance engineering

In looking back at the two Haystack events Charlie Hull (Flax) commented in a [blog post](#)

“Those of us who have been working in the search sector for a while know that search tuning isn’t just a matter of installing the default configuration, pointing the engine at some content and starting it up – in fact, if you do just that you’ll probably end up with a search user experience that’s even worse than whatever you’re replacing and certainly a lot worse than your competitors’ solution. It’s also no longer about just knowing how one engine behaves and the magic tweaks to improve it – you need to understand the fundamentals of search and how a range of different products and projects implement this. You also need to understand user requirements and their often entirely subjective views of what is a ‘good’ and ‘bad’ search result, plus how different types of businesses can use search technology for site search, enterprise search, media monitoring, process improvement and myriad of other uses.”

An outcome of this analysis of the situation is the development by Charlie Hull of the role of Relevance Engineer.

GDPR implications for cognitive search

The implementation of the EU General Data Protection Regulations took place in May 2018. The implications for search managers are only just starting to emerge. One of the most complex areas is the extent to which it is permissible to use transaction and location data from employees to provide ‘personalised’ results.

In November 2018 the [Dutch Government published a 90-page report](#) on the data protection issues with the logging software that Microsoft uses in its Office software. It had been commissioned by the Ministry of Justice and Security for the benefit of SLM Rijk

because the Government was concerned about the information that Microsoft was collecting through the logging routines built into the application.

The report notes that discussions have been held with Microsoft, but the issues are still open ones. This is not surprising as these sub-routines go to the heart of how Microsoft delivers functionality. The potential GDPR issues of employee monitoring have also been considered by the Article 29 Data Protection Working Party of the European Commission in Opinion 2/2017 on data processing at work adopted on 8 June 2017.

At present search vendors (most of whom are US-owned) are promoting the use of logging software in order to deliver personalised sets of results to employees using their search application and to identify the expertise of employees so that others can identify potential experts within their organisation. In theory this seems to be a very helpful initiative. The emphasis on 'US owned' is because few of these vendors put the GDPR implications of their technology visibly in front of potential customers.

However, there are now concerns on two fronts about their impact. The first of these is whether they satisfy the requirements of GDPR. This is the reason for the actions taken by the Dutch Ministry of Justice and Security. It is not just the raw data but the weighting that has been put on each element. The situation becomes even more challenging in an enterprise situation when the core factor in deciding on the selection of information to present is the security permissions of the employee.

The ethics of personalisation

There is a wider issue about the ethics of personalisation, and the extent to which the algorithms should be transparent. The fundamental basis of AI is to use the past to predict the future. In 2018 Amazon revealed that its AI-based recruitment application was biased towards male candidates because in the past there had been just such a bias. In business, flexibility of response to an opportunity or a challenge is essential, requiring a search application that can be used to think outside of the box and not just repeat the approaches used in the past.

A good [overview](#) of the different ways in which recommender algorithms work was published as far back as 2007. The fundamental issues of algorithm transparency have not changed since that time, but now the sophistication and abundance of algorithmic approaches to personalisation require far more careful consideration than they did more than a decade ago. In 2018 a very important [review paper](#) entitled Evaluation in Contextual Information Retrieval: Foundations and Recent Advances within the Challenges of Context Dynamicity and Data Privacy was published by Lynda Tamine and Mariam Daoud. This paper sets out the issues very clearly and indicates that so far the level of technical development in understanding the context of search has not been matched by an assessment of the implications of the technology on users and on society.

The rise of the relevance engineer

Charlie Hull and Doug Turnbull

Perhaps you've noticed a trend lately.

Search technology roles at companies have taken on a new flavour. Yes, there are the traditional roles focused around search engine technology a 'search engineer' focused on all aspects of the search engine including setting up a search engine, understanding the data structures, building search applications, improving performance, and perhaps tweaking the weights of a few fields in search.

The last few years has seen the emergence of a new role on search teams. The broader 'search engineer' role has been refined, with companies now offering roles for 'relevance engineers'. If you do a job search for this title, you'll see results like this:



Team Lead Engineering - Search & Relevance

Grubhub

New York, NY, US

Experience with Search Engines and SDK's eg Lucene, Solr, Elasticsearch etc. Familiarity or e... careers-grubhub.icims.com

Search and Relevance Engineer

Walmart eCommerce ★★★★★ 280 reviews

Sunnyvale, CA 94087

Experience with Elasticsearch or Apache Solr is preferred. As a Search & Relevance Engineer on the Sam's Club Search Team, your mission is to design, build, and...

30+ days ago [save job](#) [more...](#)

Search and Discovery Engineer

Alphasights: Engineering

New York, NY

Search and Discovery Engineer. Constantly learn from and mentor other engineers. Build elegant components that leverage our data to make discovery smarter and...

30+ days ago [save job](#) [more...](#)



Relevance Engineer

Wish

San Francisco, CA, US

Excellent software design skill with experience in Python, Go, and/or Java. The role is ideal for so... jobs.lever.co

8 months ago

(That is, you'd hope to see those titles if the job search engine's relevance was any good, which is not always the case.)

What is a relevance engineer?

What exactly is a 'relevance engineer' and how is it different from a 'search engineer'?

A search engineer focuses broadly on all of search's concerns, with relevance (maybe) one of many other concerns, along with performance and application development. A relevance engineer is a deeper specialisation focused on whether a search system answers user questions effectively. Because, as it turns out, search engines don't do a particularly good job of answering our users' questions without significant manipulation of the search engine technology.

Both performance and user experience are key to delivering the best search quality. But the relevance engineer's deeper specialisation is accurately answering the user's question. The challenges of a relevance engineer include:

- How can one determine whether a search solution is successful at user or organisational goals?
- How can one measure what the user actually means when they type in a specific search engine query?
- How does one use the organisation's knowledge assets in a search engine to manipulate ranking to meet those user goals?
- How is the search engine manipulated or tuned to meet user goals?

None of these are easy questions, and we'll explore the unique challenges of relevance engineering that every organisation faces later.

Where did the 'relevance engineer' come from?

Where did this profession come from? The story is rooted in open source.

The 2000s saw the rise of open source search, primarily based on the open source library Lucene. Out of Lucene emerged two search engines: Solr and Elasticsearch. During the late 2000s and early 2010s organisations with search needs sidestepped expensive, proprietary solutions and turned to open source search. The development also paralleled the emergence of 'NoSQL' technologies. With the NoSQL movement, developers became comfortable with data stores not based on traditional relational databases.

Setting up search became easy for most IT departments to tackle. Much of the open source search development was focused around building straightforward search applications with basic requirements. Initially organisations were satisfied enough with the simple, default relevance scoring within these search engines: either the ranking was good enough, or hardly paid attention to. At most, fields were given a few weights or 'boosts' to attempt to prioritise them.

Organisations too were undergoing a digital transformation, moving face-to-face, real-world presence to focus on online presence. Users came to a website and wanted to 'talk' to someone: they wanted to ask the same question they would ask a sales person or a librarian. Trained by Google and Amazon, they wanted search to 'get them' with very high accuracy - and getting search wrong became analogous to rude or unhelpful service. The stakes were (and still are) high.

Some search engineers began to specialise more deeply in these challenging problems, creating the speciality of relevance engineering.

They quickly found the challenge daunting. It became apparent that a given application's relevance requirements were just as unique as every other part of the application. An online shoe store's relevance solution might look nothing like that of an online book store. A shoe store needs to understand shoe sizes, colours, styles; a book store: authors, title, subjects - and all of these would look nothing like job search, enterprise search, or any number of other applications.

Users were coming in droves to these applications, but answers were not in abundance. Looking to academia didn't give much to those budding relevance engineers. Certain practices developed decades ago helped create a common set of principles, yet beyond that, very little academic research occurred beyond that performed on Web Search in the late 2000s and early 2010s. To this day, major search companies like Google and Microsoft dominate the Information Retrieval research community.

This brings us to where we are today. Relevance engineering is an emerging field ripe for innovation. What works for the book store may not work for the shoe store. Neither may work for the electronics store, or the newspaper, or the job search or dating site. Books like *Relevant Search* have explored some practices, and a community of consultants and freelancers have stepped up to meet the challenge organisations face. Conferences like [Haystack](#) help. Academic information retrieval is giving greater focus on topics beyond Web search and major academic conferences like SIGIR and ECIR have an industry track. But a tremendous amount of work must be done to fully define this field.

The relevance engineer's persistent problem: measurement

Relevance engineers work to manipulate the search engine to do their bidding. Manipulating a search engine seems hard enough. But it turns out there's an even tougher problem:

What did the user even *mean* when they typed a query like 'december projections' into the search bar?

Does this user want next December's projections? Last December's projections? Are 'projections' relative to a department or specific business the user is tied to?

Understanding the right answers for search query takes an immense amount of effort. One possible solution is to simply monitor what users click or interact with. But users only click on what they see. If the right answer is buried on page 50 of the search results they will never click on the right result.

Another solution is to work directly with users to understand what their queries mean. Dear user, what *is* the right answer to this query? And oh, by the way, what are the wrong answers? And which answers get it kind-of right? Yet a good search system has millions of queries. The bulk of queries are in the "long tail", and often obscure.

A good relevance engineer is obsessed with this challenging problem of measurement. The best teams pore over user feedback and analytics data, struggling to get any sense of whether a user was satisfied with the result. And if they were satisfied, what document ultimately scratched their itch?

Always with relevance engineering: arriving at a good solution requires more than technical expertise. Interpreting and acting on user feedback requires collaboration with product stakeholders. Understanding the context a user is operating within, for a specific application, that prompts them to type 'december projections' takes a tremendous amount of skill in the domain.

How can you grow your relevance engineering capabilities?

Despite the success of Haystack and its subsequent European counterpart held in London in October 2018, there is still a huge need for search and relevance expertise. Demand outstrips supply, with far more jobs available than applicants to fill them.

How do you hire for such an in-demand profession? The short answer is, you can't. Don't continue chasing unicorns in the vain hope that the job market will provide. The key to building your own competency is to build rather than hire. Relevance engineers aren't born but made. Instead of a 'recruiting only' strategy, companies should be encouraging and supporting their staff in acquiring relevance engineering skills. Training is available in this important new discipline. Consulting firms specialise in growing your own internal capabilities in search and relevance.

The broader community has a hand to play in shortening the gap. The information retrieval field has long focused on Web search, and not paid much attention to industry-standard open source tooling. Universities that teach information retrieval and related courses should encourage their students to gain practical experience of search tuning using industry-standard open source search engines. Graduate research work should move away from just the concerns of Web search and address other fields.

The community is also stepping up to fill the relevance engineering gap with tools and techniques.

Encouragingly we are seeing the creation of a raft of specialist tools useful for relevance engineering. OpenSource Connections' [Quepid](#) provides a browser-based relevance tuning workbench and [Sease Ltd's](#) Rated Ranking Evaluator (RRE) offers a way to run hundreds or even thousands of relevance measurements on each new search configuration. [Luigi's Box](#), winner of the Best Startup category at the British Computer Society's annual Search Solutions Awards, provides a powerful online dashboard for search queries. This is only a selection of the tools available and more are appearing all the time.

We also see methodologies and techniques to describe, measure and tune search engines being made available in blogs and conference talks. The most useful of these documents even failing strategies with refreshing honesty. This is particularly good for machine learning based approaches to search tuning, where it is becoming apparent that without a reliable and sufficient set of data, the ability to evolve understandable models or (most importantly) the right team of people, it is very easy to spend a lot of time with no useful result. These 'AI' approaches, while currently fashionable, are hard to get right and until we understand what doesn't work, we will make no real progress beyond the marketing spin.

The community has a huge job ahead of it, and is always eager for new members to share what they know. It's an exciting time to become a relevance engineer! The exploration of existing domains like e-commerce continues, and as the market expands to new uses of search, so must the relevance engineer strive to transcend what has been achieved in other fields.

Rethinking 'Advanced Search': a new approach to complex query formulation

Tony Russell-Rose and Jon Chamberlain

Introduction

Many knowledge workers rely on the effective use of search applications in the course of their professional duties [6]. Patent agents, for example, depend on accurate prior art search as the foundation of their due diligence process [10]. Similarly, recruitment professionals rely on Boolean search as the basis of the candidate sourcing process [8], and media monitoring professionals routinely manage thousands of Boolean expressions on behalf their client briefs [12].

The traditional solution is to formulate complex Boolean expressions consisting of keywords, operators and search commands, such as that shown in Figure 1. However, the practice of using Boolean strings to articulate complex information needs suffers from a number of fundamental shortcomings [9]. First, it is poor at communicating structure: without some sort of physical cue such as indentation, parentheses and other delimiters can become lost among other alphanumeric characters. Second, it scales poorly: as queries grow in size, readability becomes progressively degraded. Third, they are error-prone: even if syntax checking is provided, it is still possible to place parentheses incorrectly, changing the semantics of the whole expression.

```
(cv OR "cirriculum vitae" OR resume OR "resum") (file-  
type:doc OR filetype:pdf OR filetype:txt) (inurl:profile  
OR inurl:cv OR inurl:resume OR initile:profile OR inti-  
tile:cv OR initile:resume) ("project manager" OR "it  
project manager" OR "program* manager" OR "data migration  
manager" OR "data migration project manager") (lein-  
ster OR munster OR ulster OR connaught OR dublin)  
-template -sample -example -tutorial -builder -"writing  
tips" -apply -advert -consultancy
```

Fig. 1: An example from the Boolean Search Strings Repository

To mitigate these issues, many professionals rely on previous examples of best practice. Recruitment professionals, for example, draw on repositories such as the Boolean Search Strings Repository and the Boolean String Bank. However, these repositories store content as unstructured text strings, and as such their true value as a source of experimentation and learning may never be fully realised.

2dSearch offers an alternative approach. Instead of formulating Boolean strings, queries are expressed by combining objects on a two-dimensional canvas and relationships are articulated using direct manipulation. This eliminates many sources of syntactic error, makes the query semantics more transparent, and offers further opportunities for query refinement and optimisation.

Related work

The application of data visualisation to search query formulation can offer significant benefits, such as fewer zero-hit queries, improved query comprehension and better support for exploration of an unfamiliar database [3]. An early example is that of Anick et al. [1], who developed a two-dimensional graphical representation of a user's natural language query that supported reformulation via direct manipulation. Fishkin and

Stone [2] investigated the application of direct manipulation techniques to database query formulation, using a system of lenses to refine and filter the data. Jones [4] developed a query interface to the New Zealand Digital Library which uses Venn diagrams and integrated query result previews.

A further example is Yi et al. [13], who applied a dust and magnet metaphor to multi-variate data visualisation. Nitsche and Nurnberger [5] developed a system based on a radial user interface that supports phrasing and interactive visual refinement of vague queries. A further example is Boolify, which provides a drag and drop interface to Google. More recently, de Vries et al [11] developed a system which utilises a visual canvas and elementary building blocks to allow users to graphically configure a search engine. 2dSearch differs from the prior art in offering a database-agnostic approach with automated query suggestions and support for optimising, sharing and re-using query templates and best practices.

Design concept

At the heart of 2dSearch is a graphical editor which allows the user to formulate queries as objects on a two-dimensional canvas. Concepts can be simple keywords or attribute:value pairs representing controlled vocabulary terms or database-specific search operators. Concepts can be combined using Boolean (and other) operators to form higher-level groups and then iteratively nested to create expressions of arbitrary complexity. Groups can be expanded or collapsed on demand to facilitate transparency and readability.

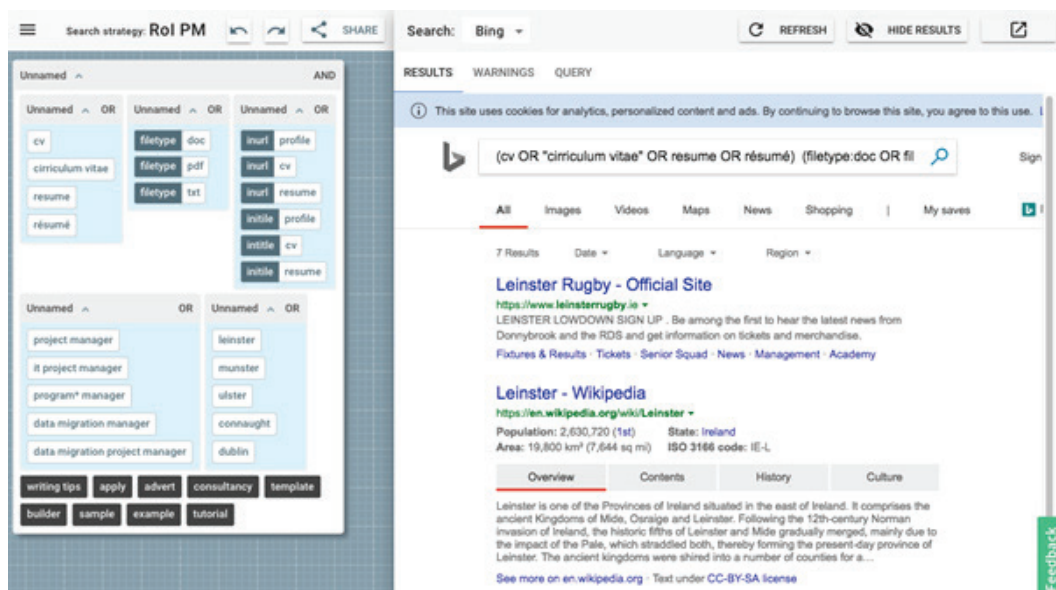


Fig. 2: The 2dSearch app showing query canvas (left) and search results pane (right)

The application consists of two panes (see Figure 2): a query canvas and a search results pane (which can be resized or detached in a separate window). The canvas can be resized or zoomed, and features an 'overview' widget to allow users to navigate to elements that may be outside the current viewport. Adopting design cues from Google's Material Design language, a sliding menu is offered on the left, providing file I/O and other options. This is complemented by a navigation bar which provides support for document-level functions such as naming and sharing queries.

Although 2dSearch supports the creation of complex queries from a blank canvas, its value is most readily understood by reference to an example such as that of Figure 1, which is intended to find social profiles for data migration project managers located in Dublin. Although relatively simple, this query is still difficult to interpret, optimise or debug. However, when opened with 2dSearch, it becomes apparent that the overall expression consists of a conjunction of OR clauses (nested blocks) with a number of specialist search operators (dark blue) and negated terms (white on black). To edit the expression, the user can move terms using direct manipulation or create new groups by combining terms. They can also cut, copy, delete, and lasso multiple objects. If they want to understand the effect of one group in isolation, they can execute it individually. Conversely, if they want to remove one element from consideration, they can disable it. In each case, the effects of each operation are displayed in real time in the adjacent search results pane.

2dSearch functions as a meta-search engine, so is in principle agnostic of any particular search technology or platform. In practice however, to execute a given query, the semantics of the canvas content must be mapped to the API of the underlying database. This is achieved via an abstraction layer or set of ‘adapters’ for common search platforms such as Bing, Google, PubMed, Google Scholar, etc. These are user selectable via a drop-down control.

Support for query optimisation is provided via a ‘Messages’ tab on the results pane. For example, if the user tries to execute via Bing a query string containing operators specific to Google, an alert is shown listing the unknown operators. 2dSearch also identifies redundant structure (e.g. spurious brackets or duplicate elements) and supports comparison of canonical representations. Query suggestions are provided via an NLP services API which utilises various Python libraries (for word embedding, keyword extraction, etc.) and SPARQL endpoints (for linked open data ontology lookup) [7].

Summary and further work

2dSearch is a framework for search query formulation in which information needs are expressed by manipulating objects on a two-dimensional canvas. Transforming logical structure into physical structure mitigates many of the shortcomings of Boolean strings. This eliminates syntax errors, makes the query semantics more transparent and offers new ways to optimise, save and share best practices. In due course, we hope to engage in a formal, user-centric evaluation, particularly in relation to traditional query builders. We are currently engaging in an outreach programme and invite subject matter experts to work with us in building repositories of curated (or user generated) examples and templates.

Adopting a database-agnostic approach presents challenges, but it also offers the prospect of a universal framework in which information needs can be articulated in a generic manner and the task of mapping to an underlying database can be delegated to platform-specific adapters. This could have profound implications for the way in which professional search skills are taught, learnt and applied.

References

1. Anick, P.G., Brennan, J.D., Flynn, R.A., Hanssen, D.R., Alvey, B., Robbins, J.M.: A direct manipulation interface for Boolean information retrieval via natural language query. In: Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 135–150. SIGIR '90, ACM, New York, NY, USA (1990). <https://doi.org/10.1145/96749.98015>
2. Fishkin, K., Stone, M.C.: Enhanced dynamic queries via movable filters. pp. 415–420. ACM Press (1995)
3. Goldberg, J.H., Gajendar, U.N.: Graphical condition builder for facilitating database queries. U.S. Patent No. 7,383,513. 3 (2008)
4. Jones, S.: Graphical query specification and dynamic result previews for a digital library. In: Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology. pp. 143–151. UIST '98, ACM, New York, NY, USA (1998). <https://doi.org/10.1145/288392.288595>
5. Nitsche, M., Nürnberger, A.: Quest: Querying complex information by direct manipulation. In: Yamamoto S. (eds) Human Interface and the Management of Information. Information and Interaction Design. HIMI 2013. Lecture Notes in Computer Science 8016 (2006)
6. Russell-Rose, T., Chamberlain, J., Azzopardi, L.: Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management* 54 (6), 1042–1057 (2018). <https://doi.org/10.1016/j.ipm.2018.07.003>
7. Russell-Rose, T., Gooch, P.: 2dsearch: A visual approach to search strategy formulation. In: Proceedings of DESIRES: Design of Experimental Search Information Retrieval Systems. DESIRES 2018 (2018)
8. Russell-Rose, T., Chamberlain, J.: Real-world expertise retrieval: The information seeking behaviour of recruitment professionals. In: European Conference on Information Retrieval. pp. 669–674. Springer (2016)
9. Russell-Rose, T., Chamberlain, J.: Searching for talent: The information retrieval challenges of recruitment professionals. *Business Information Review* 33 (1), 40–48 (2016)
10. Tait, J.I.: An introduction to professional search. In: Professional search in the modern world, pp. 1–5. Springer (2014)
11. de Vries, A.P., Alink, W., Cornacchia, R.: Search by strategy. In: Proceedings of the third workshop on Exploiting semantic annotations in information retrieval. pp. 27–28. ACM (2010)
12. Wing Pazer, J.: The importance of the Boolean search query in social media monitoring tools. DragonSearch white paper (2013), <https://www.dragon360.com/wp-content/uploads/2013/08/social-media-monitoring-tools-boolean-search-query.pdf> (retrieved 22-Mar-2018)
13. Yi, J.S., Melton, R., Stasko, J., Jacko, J.A.: Dust & magnet: Multivariate information visualization using a magnet metaphor. *Information Visualization* 4 (4), 239–256 (Oct 2005). <https://doi.org/10.1057/palgrave.ivs.9500099>

Search insights - a view from a search manager

Karen Renshaw

For any eCommerce organisation on-site search is a key element in driving a successful customer purchase. However, for many years on-site search was long ignored as being something that would just work 'out of the box' or was the remit of the technical team. This resulted in not only a below par experience for the customer but was also costly for the business, as customers unable to find the product they were looking for would go elsewhere.

I have worked within on-site search for over 12 years, but I come to it not with an academic background, degree in Information Science (or similar) but as someone who has built up their experience as they have gone, learning from implementations that haven't delivered against the business case as expected.

As a search manager responsible for driving improvements to search it was almost impossible to find a demonstrable method focused on creating, implementing and measuring a framework to drive relevancy improvements.

Very much viewed as a black art, relevancy improvements were largely ignored with search changes focusing on user interface and design changes. These were a very visual demonstration, when senior managers and stakeholders demanded 'search be fixed', but they never got to the root of addressing the core problem of ensuring that a relevant set of results was returned. After all, how can you create a set of requirements into IT asking for more a more relevant set of results (and as importantly acceptance criteria), when from a technical perspective the results are relevant as they are a match to the content being indexed?

Thankfully, the world is changing, and on-site search is emerging as a discipline in its own right with more and more resources available to help search managers understand how to best approach delivering a better search experience.

Here, I am focusing on how to get started on improving the search experience from a relevancy perspective. Relevancy of results should be considered within the context of the overall search experience.

A good search experience is made up of:

1. Relevant results - results that make sense against the query entered
2. UI that supports the customer need to navigate around the results set
3. Clear messaging that communicates the actions that have been taken
4. Consistent, normalised content

All of these elements must be considered together when making search changes. Depending on organisation structure, different teams may have responsibility for each element, but it is key to ensure that the customer can find, select and buy the product they need easily.

There is a framework which I have used to implement a series of relevancy changes - more details below - but there are two key messages that I would emphasise above all else.

1. Search is not a plug and play situation. It is not a case of replacing one technology with another, assuming that the default configuration will work, and that the benefits will automatically flow through. As with any change, success is driven through the combination of people, processes and technology. Only by understanding what you want to achieve, can you harness the technology to deliver the results that you are looking for.
2. Search is not a one-off activity. Making search work for your organisation requires ongoing, iterative review and changes. Even with a good implementation the way in which customers search, the content and the product set will change and as such that will change the results displayed to customers. A constant review of how search queries are performing and what changes you can make to improve them should be undertaken. And, that requires dedicated support within your business, from someone who understands the business objectives, how search works and more importantly the desired customer experience.

What is a search team responsible for?

There are a number of ways that search teams can optimise the search experience, some of which they will have responsibility for delivery of and others they should be closely aligned with. The key elements are:

- 1) Review high volume / underperforming search queries and manage manually
- 2) Improving and testing the core configuration (algorithms)
- 3) Managing the navigation path
- 4) Defining content changes

Reviewing and improving top search queries is the easiest way to get started on driving improvements, as typically these queries can be improved through the addition of a synonym or content update. These types of changes can drive incremental conversion uplifts so shouldn't be ignored.

Past these high-volume queries, investment should be made in improving the long tail through ensuring that the search set up is configured to meet the needs of your customers. The search set up will be different by organisation, as it will need to be tailored to content and customer behaviour.

Whilst this can seem daunting creating a relevancy framework by which you can identify, test and measure changes helps you to get started.

The framework

1) Understand your customers and how they search

- This is critical. Knowing how your customers search not only helps you build out a customer driven relevancy strategy, but also ensures that you build out the overall experience that supports the customers' ability to find, choose and select the right product.
- Similarly, if there is a lot of organisational noise around your search experience it can be easy to get distracted to 'fix' specific 'visible (but as important) queries' and to tune the engine to deal with these, but this approach can lead to other less 'visible (but as important) queries'
- Being aware of how your customers search can then help you to 'navigate' those internal conversations.

2) Understand the current set up - both search engine and content

- Having a base understanding of how your search is currently configured will help you to start to diagnose any issues and the changes that you might want to make.
- Knowing which content fields are indexed, how much weighting is applied to each, how specific or loose the engine is set up to return results will allow you to decide which elements to test.

3) Understand how well your search queries are performing today

- In order to be confident in the results that you are driving, and to measure the changes that you are expecting, before embarking on any change, create a series of benchmark query sets.
- The purpose being that these can be measured throughout the testing, providing a view of the impact changes are having.

4) Create a test matrix

- There are multiple ways that you can configure search to improve underperforming sets of queries. Knowing which one will work for your business and drive the change you want will require testing.
- Plan for incremental changes.

5) Test, test, test - then test some more!

- Only when changes are deployed against a full data set can the real impact of the change be seen. Sometimes the configuration might not drive out the results that are expected. Having a defined set of queries that you can test against will help to ensure that changes are working as expected.
- It's always worth testing queries that you didn't think would be impacted by the changes too, so that you can ensure that you haven't created another issue.
- The amount of testing you choose to do is purely a business decision, but I would always suggest a minimum amount before releasing into production (even if you are A/B testing) so that you can identify anything that could impact conversion.

This approach provides a baseline against which on-going search improvements can be made. But to link back to the earlier key messages it is just that - a baseline - as search technology continues to change so will customer expectations and that means adopting an iterative review of how relevant your search is!

Microsoft Search

Agnes Molnar

In recent years, Search in SharePoint has been in a kind of ‘sedentary’ mode. No new features or functional improvements have been rolled out. Many people were saying that SharePoint Search was dead.

Over the last few months there have been signs that we should expect changes to Office 365. Microsoft Graph has got stronger, and Microsoft’s ‘cloud first’ strategy suggests that any future changes and developments would be in the cloud.

The evolution of Microsoft Enterprise Search

A brief historical overview will help us understand the current situation.

In 2008, Microsoft acquired FAST Search & Transfer, a market leading enterprise search company headquartered in Oslo, Norway. The goal was to integrate FAST Search into SharePoint to provide cutting-edge, intelligent search capabilities.

For the next version, SharePoint 2010, the integration was not fully complete. Instead, SharePoint 2010 Server had its ‘own’ search engine, based on the old SharePoint 2007 Server technology, although with major improvements. Companies could also decide to purchase FAST Search for SharePoint 2010 (FS4SP) as a separate product. It had to be licensed separately and installed on a separate server but could be integrated with SharePoint 2010 Server (Enterprise license only), and provide additional, enhanced search capabilities, based on the FAST technology.

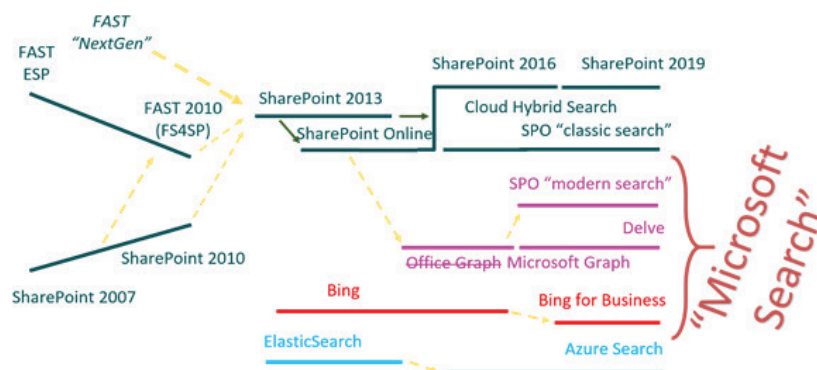
The integration was completed for the next version, SharePoint 2013. Instead of ‘FAST Search for SharePoint’ being a separately licensed product, everyone who installed SharePoint got the ‘new’ search engine.

At the same time, SharePoint Online was born in the cloud. Based on SharePoint 2013’s source code, it also featured SharePoint 2013’s search engine.

This was also the last time Microsoft added any functionality to ‘classic’ search. Instead, Microsoft has started to invest huge amounts of time, money and resources into Microsoft Graph – the technology that has been driving every product release decision. If you use Office 365, you cannot avoid learning about Microsoft Graph.

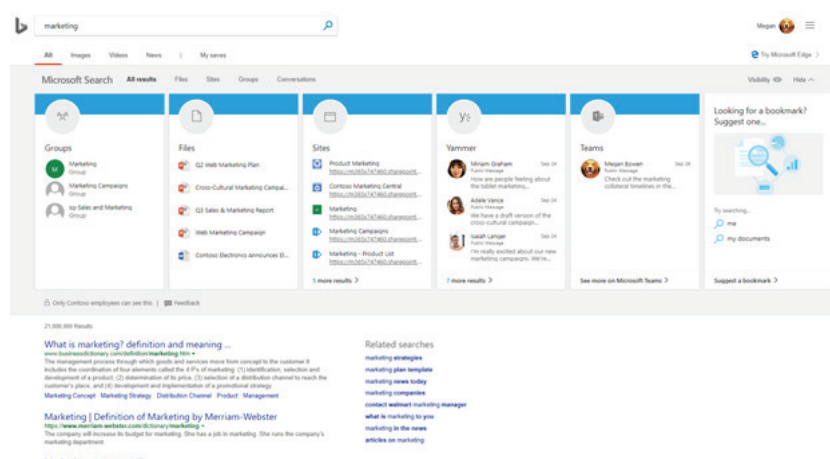
SharePoint 2016 and 2019 did not bring any new features or major enhancements to ‘classic’ search. Neither did Office 365. ‘Classic’ search has remained the same for the last five years.

However, ‘modern’ search was born in Office 365, using Microsoft Graph instead of the classic search index. You can find this personalised search experience everywhere in Office 365 - SharePoint Online, Exchange Online, OneDrive, Office365.com, Delve, Teams. The problem is that although each of these experiences uses Microsoft Graph, the user experience is different everywhere.



History of Microsoft Search (credit: Jeff Fried)

At the same time, Bing has evolved into one of the Google's primary competitors in public web search. At Microsoft Ignite 2017, Microsoft also announced a new search technology called 'Bing for Business', which provides enterprise search results embedded into web searches if the user is logged in with his/her Office 365 profile to Bing.



Enterprise Search results embedded into web search results in Bing

Microsoft also offers Azure Search. This is search-as-a-service based on Elastic Search.

Office 365 Search today

The 'classic' SharePoint Online Search provides more or less the same features as its on-prem version in SharePoint 2013, 2016 and 2019. The truth of the matter is that there have been no changes since SharePoint 2013!

However, because it can be customised and configured, SharePoint Online 'Classic' search remains the primary choice of the majority of organisations. Even in late 2018, we still do not have any real control over what results the users can see in 'modern' search, how these results are ranked, or how they are being displayed: no custom display templates, no custom facets (refiners), no custom search verticals. We have to rely on the out-of-the-box, mostly black box algorithms, driven by Microsoft Graph. 'Personalisation' of the results also makes it hard to understand and explain Search to end users, and this leads to more and more confusion.

The breakthrough happened at the Microsoft Ignite conference in September 2017. In his conference keynote Satya Nadella, CEO of Microsoft, emphasised the importance

of search and shared the vision of a new wave of modern, personalised search in Office 365. It was a surprise that Search was a vital topic in his conference keynote – this had not happened since the acquisition of FAST Search and Transfer, almost ten years before.

Yet this was just the beginning. 2018 was the year of big search announcements at two major Microsoft conferences: SharePoint Conference North America in May, followed by Microsoft Ignite in September.

Announcing Microsoft Search

While the potential of Graph-driven, intelligent and ‘personalised’ search is clear, there are still many open questions. After years of discussions with industry experts and enterprise customers, Microsoft finally concluded that a wave of significant improvements was needed. Microsoft Graph is already mature enough to support significant search upgrades.

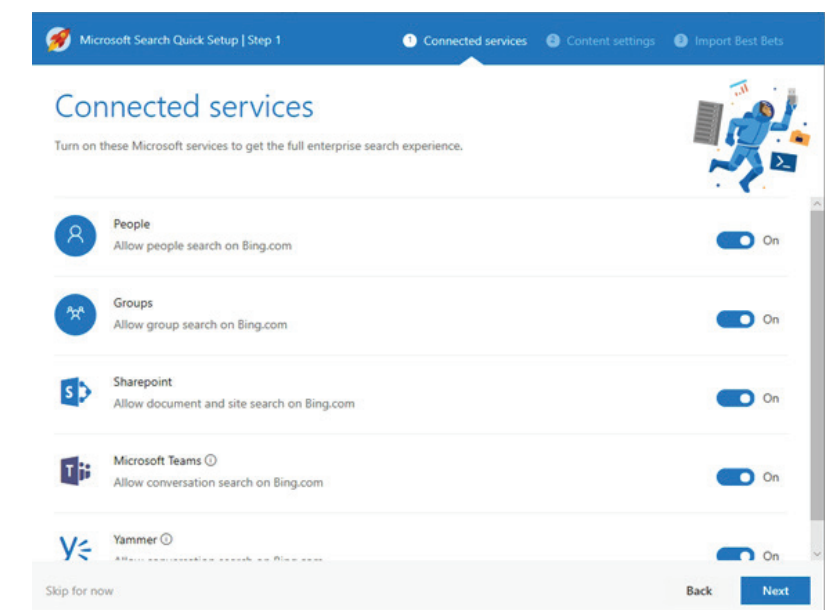
But first, an important decision had to be made. Office 365 has many different applications with different user search experiences. In 2018 Microsoft publicly announced its commitment to improving the modern search experience in Office 365. This consolidation also includes Bing for Business, which is going to be the part of the new Microsoft Search family.

At the time of writing there are some things we already know about Microsoft Search, some things we can take an educated guess at, and some things we simply don’t know yet.

The following summary is my current view on what we can expect. But remember, Office 365 is in the cloud, and things can and do change on a weekly or even a daily basis. I highly recommend you follow my blog at <https://SearchExplained.com/blog> for the very latest updates.

What we know about Microsoft Search so far, and what we can expect?

- Office 365 administrators can enable Microsoft Search if the tenant is hosted in one of the following countries: Australia, Canada, France, Germany, India, United Kingdom, USA. After enabling Microsoft Search, various configuration options will be available under Tenant Administration / Microsoft Search admin.



- Consistent place and experience across applications: Microsoft Search comes not only in the browser but also in Office client applications (Word, PowerPoint, Excel, etc.). The location and experience of search will be consistent across all of these applications.
- Personalised results everywhere: Microsoft Search keeps providing Graph-driven, personalised results to everyone, everywhere.
- Customisation of 'modern' search, including custom facets (which are called 'filters' now, rather than 'refiners' as in 'classic' search), custom search verticals as well as the ability to use SharePoint Framework (SPFx) to define how the results should be displayed (instead of the 'classic' display templates).
- Unified administration: administration of Microsoft Search will be centralised into the Office 365 Tenant Administration, to provide consistency here as well.
- Curated results and bookmarks: 'Best bets' can be defined to help users with centrally curated search results.
- Third party connectors and APIs: Microsoft has also announced that it will develop third-party connectors as well as search connector APIs. Details are still rather fuzzy. We don't know any public information yet about what connectors they are working on or when and how the connector APIs can be used. It's important to mention though, that these connectors and APIs will be available in Office 365's Microsoft Search (the current connector APIs are available only on-prem today).

Conclusions

Preparing an organisation for a new Search application has never been easy. Office 365 offers tremendous and compelling features with Microsoft Search. However it also adds complexity to user adoption and as always with Microsoft we only know what is actually on the roadmap when it arrives.

First of all, Microsoft Search results are always personalised. Ranking results is black box and there are no real options to systematically improve relevance. This also leads us to severe user adoption issues. We have to educate users about what to expect and explaining how the Microsoft Graph works is far from straightforward. Also, organisations will need to provide ongoing help and support when in general they have not invested in the skills and experience needed to get the best out of the complex array of Microsoft offerings.

Based on my experience you will need to have good quality, trustworthy external help with these daunting adoption challenges as Microsoft does very little to support a search team in delivering an optimal search experience to their organisation.

Why we (still) need taxonomies (and the taxonomists who nurture them)

Helen Lippell

In late 2018, for the third year running, I had the tremendous privilege of being Programme Chair of [Taxonomy Boot Camp London](#) (TBCL). This event brings together a diverse community of people who manage or use taxonomies in myriad applications for research, publishing, navigation, data analysis, and of course, search.

One of the main themes of TBCL 2018 (and of its North American [sister event](#) in Washington DC) was the excitement about the potential of using AI and machine learning to enhance data-driven applications. There are ambitious claims by some solution vendors about what is possible. There are perennial news articles about ‘robots replacing all our jobs’.

Yet practitioners have always adapted to technical innovation and flourished as they learn to make the most of new technical capabilities. It is therefore worth reflecting on how we as a community can add value. Here, I hope to move beyond the buzzwords to highlight the most important areas where we can make a real difference in our organisations and wider communities.

Human communication is full of nuance, tacit knowledge, and randomness

One of the key uses of a taxonomy, in its original sense from library science, is as a controlled vocabulary within a certain domain. The value of this control is in supporting consistent classification of resources and building a shared understanding of the language needed by the taxonomy’s users to complete tasks.

A taxonomy gives enterprise search relevancy algorithms a ‘leg-up’ in deciding which content may be relevant (or interesting) in response to the search user’s query. It can provide facets and filters for users to refine the initial query. A taxonomy which makes use of synonyms is even more powerful. We know that search users, whether they’re using web search or internal applications, aren’t always neat, precise or accurate when they put text into that little white search box. They never will be. It’s our job to use the tools at our disposal to best understand the content we’re working with, and the ways that users try to articulate their information needs.

The best of breed tools on the market are better than ever at processing text, automatically generating candidate concepts and taxonomies, and pulling in linked data and knowledge graph nodes. This is no existential threat to taxonomists. It’s a good thing to have a computer do the legwork of corpus analysis, data analysis and data linking, at a scale that could never be achieved manually. But merely doing these ‘word-harvesting’ jobs isn’t the whole picture. The value of human expertise is in refining, selecting, investigating and making the right connections between concepts, labels and entities. For example, an automated system might spot a link between the concepts ‘manic depression’ and ‘bipolar disorder’. A taxonomist could look into this and ascertain from experience and discussion that the former is now considered an unacceptable phrase. They might then remove the outdated phrase altogether from the taxonomy or include it only as a hidden term so that any searches for it will be steered towards the accepted term.

Scaleable knowledge models benefit from taxonomies

Ontologies are models built on defined classes of things, named relationships between those things and attributes (metadata) about those things. Knowledge graphs have become much more widely understood over the last year, and many organisations are considering implementing one. They are a superb way of modelling complex, heterogeneous domains. They can scale up to include thousands, even millions of facts.

Ontologies and knowledge graphs need taxonomies to define, classify, add colour to, and expose real-life information about things in the models. Otherwise a model is merely an extreme abstraction, a bunch of circles and lines on a screen or piece of paper. Taxonomies act as the link between the model, and the people and systems that the model serves. They are human-understandable when used for navigation, or for search facets. Everyday users should not need to engage directly with an ontology or graph. An analogy is that one does not need to understand exactly how a car engine works in order to drive.

Data is inherently biased

Despite the appearance of objectivity, data is in fact influenced by how it is defined, captured and used. Just as the map is not the territory, so data is not reality. When decisions are made about how to manage and classify data, then there is a reasonable chance that biases will creep in. The presence of bias in information ranges from the extreme (racist classifications in apartheid South Africa), to bad practice (accessibility of a venue in a listing being described as a binary yes/no without further detail for a range of accessibility needs) to the plain irritating (preponderance of pink in the design of menstruation apps).

Taxonomy is one way for organisations to be transparent about the classification and labelling choices they have made. The contents of a taxonomy should reflect the organisation's values (regardless of whether the taxonomy is for internal use or public). Taxonomists act as mediators between technology, content, business stakeholders and users/customers. They persuade, analyse, connect and align. It's a natural dimension for them to consider the wider ethical and social dimensions of their work too. Data ethics is a field that has been gaining traction recently, as technology becomes ever more pervasive in people's life, and everyone should be paying attention.

What should the future look like?

The most successful knowledge applications, now and in the future, will bring together the best aspects of knowledge engineering including taxonomies, ontologies, linked data and knowledge graphs. Likewise, enterprise search has become smarter, and the emerging field of relevance engineering will accelerate this trend. Relevance engineering will bring together technologists and practitioners like taxonomists applying their skills to make search more intuitive and useful.

Automation, artificial intelligence and innovative technologies are changing the future. But they must be led by the highest quality human oversight. Information professionals, as a core part of their role, must advocate for deep understanding of data being managed, and for ethical principles to be applied in choosing categories. In the financial sector, categorisation of customers, especially by opaque algorithms, affects which products people can access. People may be disadvantaged because of factors they are not even aware of. Scrutiny of these processes should not be left to under-resourced regulators, or to the tech sector themselves.

My conclusion therefore is that information professionals have a huge part to play in shaping these emerging and maturing technologies to serve us (and wider society) properly. Search engines in particular are often a primary means of accessing information. If the search function of a financial comparison website restricts product options for a user just because they have a 'risky' postcode, then the social impacts of this should be called out and investigated.

And as for those pesky job-stealing robots? Most CEOs would say their job is to drive business value by delivering for their customers. Information professionals use their skills to do these things too. CEOs are not going to automate themselves into irrelevance any time soon, so why should we? We have much work to do.

The impact of corporate language policies on enterprise search

Stephanie Segura Rodas

Many companies have a de facto language policy stating that English is the corporate language. It may be the default corporate language, but in any multinational company many different languages will be used. The impact of language on business operations is now the subject of many research studies. This is because language may influence not only business collaboration but also the creation of knowledge in the company. However, in the case of search application implementation, the importance of being able to search across content in a range of languages is often under-appreciated.

The scale of the challenge

Knowledge sharing is a language-based activity and exchanging information within multilingual teams brings with it the possibility of mis-communication. Having worked in multinational companies where English, French and Spanish were corporate languages, I have noticed that the results of an English-language search can be different from the results of a search enquiry in my native language (Spanish).

Globalisation and international expansion have led to a growing requirement to store and search for information in multiple languages. The table below shows the percentage of documents by language held by a major pharmaceutical company with its headquarters in Germany. The total number of content items is close to 100 million.

Language	Content items as % of total	% speaking the language as their primary language
English	73	24
German	13	25
Spanish	4	11
Portuguese	3	4
Japanese	2	6
Italian	2	5
French	1	6
Chinese	1	4
Polish	1	3

The right-hand column does not total 100% as there are a significant number of employees in Arabic-speaking countries, India and the Nordic region where English fluency is very high but English is not a national language.

There are two implications of this language diversity. The first is that although the number of content items in (say) Portuguese is small as a percentage, the importance of these content items to the Brazilian subsidiary of the company is very high. The second is that although the majority of the content items are in English, only around 24% of the 120,000 strong workforce have English as their mother language. As a result, most of the searches are carried out by employees working in their second language. A German national may have a fluent command of spoken English but may not have a good enough vocabulary to construct the optimum query.

Despite the importance of language diversity in multinational companies, there is a paucity of evidence on the impact of finding information in multiple languages. Nevertheless, it is interesting to note that even though multinational companies recognise language diversity in their communications, the uses of language diversity are only considered when a global announcement or survey is about to be released. A possible explanation for this may be the lack of resources to analyse and implement a language management policy for the organisation.

In 2018 I undertook the first ever study of the potential impact of language policies and corporate language for multinational companies in the context of Enterprise Search. This study was undertaken as the dissertation for an MBA at the Management School, University of Sheffield, and was under the direction of Professor Elaine Toms. This dissertation has provided the first comprehensive assessment of the relationship between corporate languages, searching for information in multiple languages and how this impacts work performance.

Research scope and methodology

Existing research recognises the critical role played by language policies in communication and knowledge sharing in organisations. Nevertheless, no previous research had been carried out on the impact to which the specification and operation of enterprise search applications were aligned with corporate language policies. My research set out to investigate:

- What are the reasons for the selection of the corporate language?
- What is the current state of adoption of multiple languages in enterprise applications?
- What are the issues caused by the use of multiple languages in searching, and retrieving information?

I undertook ten in-depth interviews with search managers in organisations in Europe, Asia and North America where it was accepted that employees would be working in local languages as well as the default corporate language of English. In most of these organisations I was also able to talk to communications managers who were implementing policies on language diversity.

Outcomes of the research

In this section I have presented some of the comments made by the managers I interviewed.

"I have been in the company for about 27 years, for those 27 years, the group language has been America[n] English. ... This means everything must be in English. But we also respect ... local legislation [Several countries have laws that require documentation that concerns employees to be in the native language.]"

"Companies never mention [having] an Enterprise Search according to the language policies"

"The organisation does not align very well with language diversity [...] Probably this is a result of the people procuring technology focusing [only on] English"

"At the beginning the survey was in local languages, then we changed to six [languages] and then we came up with 28 languages. We realise that employees want to express [feedback and suggestions for improvements] in their own native languages"

"IT worry [more] about elements of functionality at the system level than about the satisfaction that user[s] will gain because they do not have [a] concept of what search is"

all about from a user perspective”

“There is a team who maintain the search engine for the enterprise. They run queries but just the basic ones, I do not think they assess language-related issues”

“The reason why companies do not invest in a search analyst [is] because companies do not look [at] information as an asset. They have policies for communication, security, confidentiality, etc but companies do not have policies for information. No one owns information. So, there is no one to say if there is good or poor access to information”

“If I use the term ‘language diversity’ to an IT manager they won’t understand what I am talking about. [...] no one has the responsibility to make sure the information is good quality, if the information can be found, if it can be trusted, if it is in the right languages.”

Conclusions

This was a small-scale study of ten organisations working in a range of business sectors with a substantial global workforce. The consistency of the comments made in the interviews was quite high and it would be reasonable to assume that the outcomes would be broadly similar in many multinational organisations.

The reason for choosing English as a corporate language is that it is the most frequently spoken language in international business. However, subsidiaries use local languages for internal operations. This is because local legislation might prohibit multinational companies from using the corporate language (if this is different from the national language) for documents such as contracts and documents related to corporate compliance. As a result, MNC subsidiaries either use the corporate language set by headquarters or simply do not have a corporate language even though international expansion has increased the volume of documents in languages other than English.

Perhaps the most important part of language diversity is how this affects employees’ work performance. Researchers have shown that business decision-making depends on relevant and useful information. In addition, this research has identified that searching in various languages has become a new way of obtaining information to fulfil work assignments. Searching for information in multiple languages may have an impact on work tasks.

The results of this investigation show that little attention is paid to the use and management of information in multiple languages in multinational companies. An implication of this is the possibility that miscommunication and misunderstanding may arise within and between teams. Additionally, there may be a lack of resources to analyse and develop a language management policy for the organisation. In conclusion, the findings of this research provide insights into the effects of finding information in multiple languages in business decision-making.

Additionally, despite the challenge of searching for and retrieving information in multiple languages, multinational companies do not consider language diversity in their enterprise search applications. The outcomes of the interviews show that information stored in multiple languages is spread across the organisation without an analysis of the impact on business performance. Furthermore, this investigation has identified that multinational companies do not take into account the need for an Enterprise Search professional to implement and resolve search issues because of their insufficient understanding of search.

The outcomes of my research suggest that it is important for multinational companies to take language issues into account when considering search implementation.

Enterprise search failure

Sam Marshall

When I carry out employee focus groups for clients on their digital workplaces, it takes around three minutes before somebody complains how awful the search is. Even if that isn't what I asked about. Everyone grumbles and empathises, and the conclusion is invariably 'it should just work like Google'.

Trust me, it will never work like Google. For an intranet or digital workplace manager, it is tempting to blame the search engine or feel it is something for IT to solve. Trust me, that will never work either.

What I want to share here is a diagnostic tool that breaks down the underlying causes of search failure, and point out the many elements that intranet managers, content owners, knowledge managers, and even IT professionals can improve without changing the search engine. New research by Cleverley and Burnett attributes 62% of enterprise search dissatisfaction to non-technical factors: information quality and search literacy.

I don't want to give the impression that you shouldn't pay attention to the search engine too, but I know for many organisations search expertise can be hard to find so people end up doing nothing.

Searching step by step

Being more precise, we should talk about findability rather than search. This is because search is often a combination of searching and browsing. For example, a user might navigate to the HR section and then do a search just within that sub-site, or search for 'policies' and then navigate to 'HR policies' in a policies centre.

We can break down the search process into four basic steps:

1. Content is published
2. The search engine indexes it
3. A query retrieves a selection from the content
4. The user uses the query to complete their search

This greatly simplifies what really happens, but from a diagnostic point of view it gives us four useful starting points for things that might go wrong.

Using the tool

For each step in the process, there are things that need to go right, such as metadata, security settings and results presentation (see column 3 in Figure 1 and then underlying symptoms (the last 2 columns). Note all the ones that aren't coloured green (i.e. not primarily a technical issue)!

It's not practical to go through the diagnostic for all the content in your digital workplace. Instead what I suggest is that when you get feedback that "search isn't working", use the tool to check for systemic issues that might broadly apply to sets of content. Often, I see employee satisfaction surveys that rate search poorly, and I use focus groups to dig deeper into what's happening: "Can you remember a recent time you tried to search for something? How did you search? Did it exist at all?"

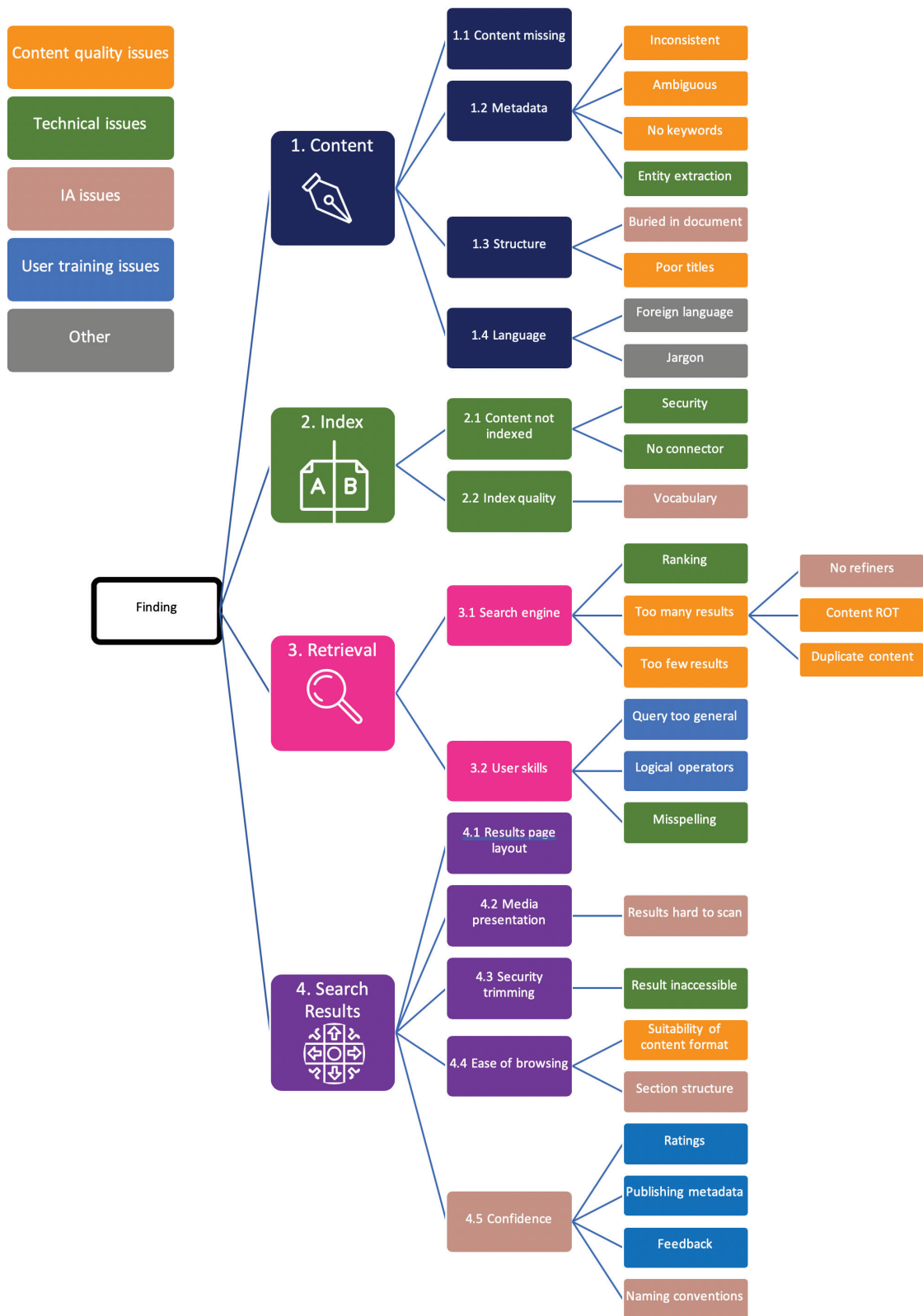


Figure 1: An enterprise search diagnostic

1. Failures of content

It sounds obvious, but often the big issue in digital workplace search is that the thing somebody is searching for just doesn't exist [1.1]. On the web, somebody, somewhere probably has put the answer there, but in the enterprise this isn't necessarily true. So if something is asked a lot, the solution might just be to get someone to write the answer

(there's a diagnostic tool for that too: [Clear Knowledge Management Roadblocks](#)). Metadata [1.2] can often be poor or lacking. Just using [good writing principles](#) for headlines and subheads can help, as can clear filenames (if you ever shared a document called "Proposal draft" or "Announcement" then I'm looking at you).

Language [1.3] can also present a barrier. A technical document may be written in jargon ('variable performance related pay') when a user searches in plain English ('bonus'). Even harder, we may expect everything to be in our language and [overlook other languages](#) ('2016 sales results for Spain' wouldn't necessarily find a document called 'Resultados de ventas de Espana 2016').

2. Indexing failures

Search retrieval works so quickly because a crawler creates an index first, and your query is actually run against the index. So, the first failure point here [2.1] is that the content needed isn't indexed. Unlike the web, a great deal of enterprise content might have security controls in place, blocking the indexer from seeing it.

More fundamentally, it may exist in a system that the crawler can't access, such as a network drive or an application. I sometimes see HR departments move all their guidelines into an employee self-service system, but if there is no connector with the enterprise search engine then routine content like 'Parental leave policy' won't get indexed nor will all those documents in dropbox as it is only shadow IT.

Next, we need to consider the index itself [2.2]. This is definitely in the technical realm but do check that document content is indexed and not just the title. You may also need to define words that are specifically meaningful to your organisation. For example, if you have a product called 'Teams', then the indexer needs to know it is more significant than casual usages of 'teams'.

3. Retrieval failures

Largely we rely on the search engine technology to get this right [3.1] and do all the good stuff like sensible ranking and knowing that 'bicycle' and 'bike' are the same. Martin White has a useful summary of [10 options for enhancing search engines](#).

However, too many results can be a symptom of duplicate content or ROT (Redundant, Outdated, Trivial), meaning a clean-up is in order. It may also mean we don't have good refiners, to whittle down results to the last six months, or only show sales collateral (see Metadata [1.2]).

Retrieval also relies on user search skills though. Google is so good we've got lazy. But enterprise search sometimes needs very good search skills, such as the use of logical operators (AND, OR, NOT). If that's unrealistic, consider [ready-made search interfaces](#).

4. Search results

Finally, we get to the results page.

You'd think if the answer was on the page we'd be successful, but if you've ever done [observational user testing](#) you'll know that sometimes people seem to fly straight past the answer and onto the phone. The layout of the results page matters [4.1], and the good news is you can often change it. Usually, the more like Google, the better, as this is what people have already learned.

Make it so that the format matches the results [4.2]: show images and videos as thumbnails, people as a contact card and, heck, even just [show the answer itself rather than a link](#).

Hits on documents can make scanning of the results harder [4.4]. If the answer is on page 52 of a document, consider breaking it into HTML pages. If the document exists but isn't shown, ask if the security settings on it are right [4.3].

Finally, users may find the right result, but carry on searching because they don't trust it [4.5]. Governance and training can help here – make sure it has things like owner and expiry details. Ratings and feedback can help too.

Credits

This post was partly inspired by an old LinkedIn thread, which Paul Culmsee analysed in forensic detail on [CleverWorkarounds](#). This is an updated version of an article [first published at CMSWire](#).

Content audit tools

David Hobbs

The pricing model for enterprise search applications is often based on the number of documents to be indexed. The process of indexing is a very substantial challenge in the implementation of a search application. Just as an illustration assume that the search application can index one document a second. A million documents will then take 11 days to index. Any index run that lasts longer than two weeks is going to be a very risky operation because if there is a major index problem along the way (such as documents in a very specialised file format) then the operation may need to be restarted from the beginning.

There are of course ways of improving indexing speed but gaining a reasonably accurate estimation of document volumes is important both from license cost and implementation perspectives. It is also an important indication of the size of the index. This might be anything from 30 to 80+% of the base volume of content. Although IT managers will have some indication of the volume of information in repositories this is almost always not a good indication of the scale of the indexing challenge. The overall challenge is to migrate high-quality information that will be of continuing value to the organisation.

Our objectives

Before we dive into tools, we should be clear about our objectives. There are two primary objectives of a content audit:

- Explore and understand our content
- Search for patterns in order to make decisions about our content

Our objective is not to stare at a list of our content. But that's what many content audits devolve into. Let's consider a hierarchy of content audit effectiveness:

- A list of content (ineffective)
- Graphs to inform decisions
- A dynamic audit that facilitates exploration (most effective)

Tools

What tools can you use for your content audit? There are four core types of tools that help in a content audit:

- Spreadsheet
- Spider
- Graphing
- Extract Transform Load (ETL)

Hint: no single tool currently available can do all of these.

Spreadsheet

Perhaps the most-used tool in content analysis is the spreadsheet, with the two most common tools being Microsoft Excel and Google Sheets. I think there are three primary reasons people use spreadsheets:

- You probably already have a spreadsheet tool
- You already know how to use it
- It's really easy to modify to suit your needs

But there are many disadvantages to a spreadsheet:

- Data, formulas, and reports are all mixed together. This makes it easy to work quickly, but it makes it easy to mess things up
- Related, although it's easier to get started it's difficult to maintain spreadsheets
- It's not natural for showing relationships and having controlled lists
- Applying the same techniques across clients/projects is error-prone

One improvement to a spreadsheet can be to use a database, which resolves all the above issues (at the expense of being more to set up in the first place). But if all you're using it for is a list of all your content then at the core you still are missing out on a stronger content audit.

Note: even with all the disadvantages of a spreadsheet, I'm not saying to never use a spreadsheet. It's just not what should be the default.

Spider

A spider methodically and automatically follows all the links on your site in order to find and discover information about the pages on your site. The primary output of a spider is fundamentally a spreadsheet of some sort listing URLs. Examples of spiders are ScreamingFrog, Xenu, and DeepCrawl.

The primary alternate method of getting a list of your content is exporting from your source system, be it a CMS, the filesystem, or a database.

There are several advantages of a spider:

- Using one does not usually require technical intervention or access
- In many ways you get to "see" the content like a visitor does (as opposed to a database dump for example, which could show content that's impossible for a visitor to even see)
- You can capture information such as link relationships

Note that most spiders are optimised for SEO analysis and not exploring and making content decisions.

Graphing tool

Graphing tools summarise your raw data in chart form.

Graphing tools are useful since they:

- Provide a view that can be shared and understood by executives and others that aren't focused on content
- Allow the content strategist or someone else content-focused to see patterns and compare buckets of content

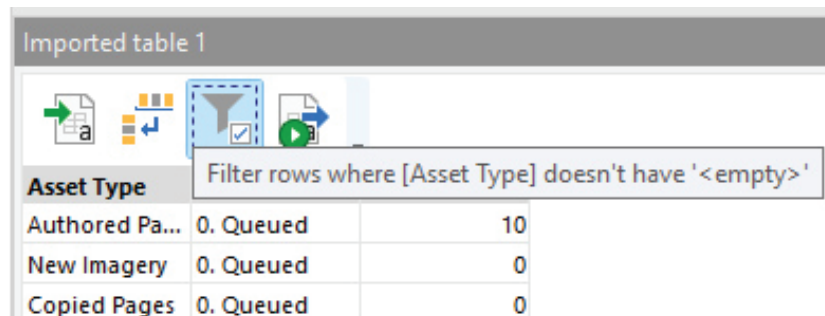
Graphing tools include Google Data Studio, Microsoft PowerBI, Tableau, and Zoho Reports. For the purposes of content analysis, the primary advantage of true graphing tools (rather than, for example, graphing in Excel) is the ability to drill down on the graphs. For instance, if you had a bar graph of the site sections and the amount of content in each, then if you clicked on "blog" bar then you could see the list of blog content.

One issue with graphing tools for content exploration is that they need to be customised to be really useful for content. For instance, when exploring content, it's usually more useful to see the biggest bars in a bar graph rather than all the bars (since you usually need to make decisions about the biggest chunks of content first).

ETL tools

Sometimes you need to massage the data in your inventory. For instance, an ETL tool could be used to merge a ScreamingFrog crawl with Google Analytics pageviews.

Examples of ETL tools include Pentaho Kettle and EasyMorph. Most ETL tools allow visual programming, like this example from EasyMorph:



The screenshot shows the EasyMorph interface. At the top, there's a header 'Imported table 1'. Below it, there are several icons representing different data operations. A filter rule is applied to the 'Asset Type' column, with the text 'Filter rows where [Asset Type] doesn't have '<empty>''. Below the filter rule, there is a table with three rows of data.

Asset Type	0. Queued	
Authored Pa...	0. Queued	10
New Imagery	0. Queued	0
Copied Pages	0. Queued	0

Example from EasyMorph

The advantages of ETL tools include:

- It's easier than coding (usually visual programming rather than traditional programming)
- Many ETL tools also will in effect aid you in debugging what's going on
- Much more repeatable processes than staying just in a spreadsheet

The primary disadvantage is that it's another type of tool to learn. Implementing things in an ETL tool is doing work that's not really at the content analysis level (even though it may be required to do the analysis, it's dropping to a fairly technical level to do so). Although it's definitely a step up from doing transformations in a spreadsheet, it still can be error-prone.

Note that you can accomplish many of these things by actual coding as well, at the expense of requiring even more expertise to do so.

Matching tools to our objectives

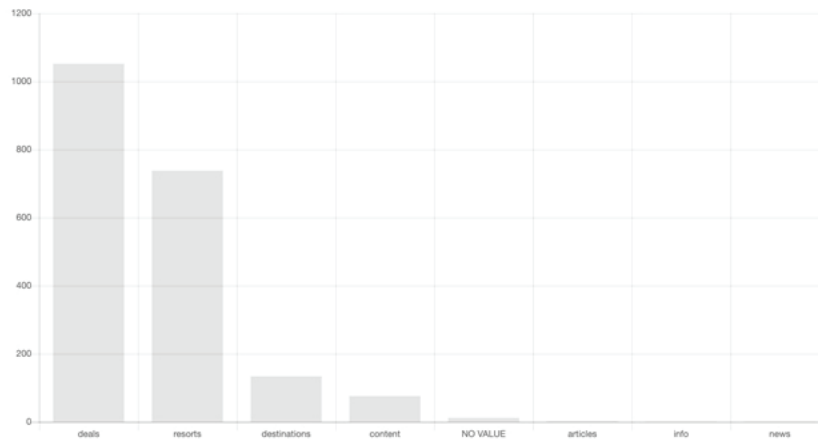
Spider and Spreadsheet → A list of content

The most popular, yet least effective, combination for content auditing is the spider and spreadsheet. Yet this just takes us to a list of content.

Spider, Spreadsheet, and Graphs → Graphs to inform decisions

An easy step forward is to use a graphing package to summarise your content in order to see patterns.

The most straightforward thing to graph is the site sections (which can often be done by taking the "folder" from the URL, such as deals in <https://davidhobbsconsulting.com/webinars/making-big-content-changes>



Example from Content Chimera

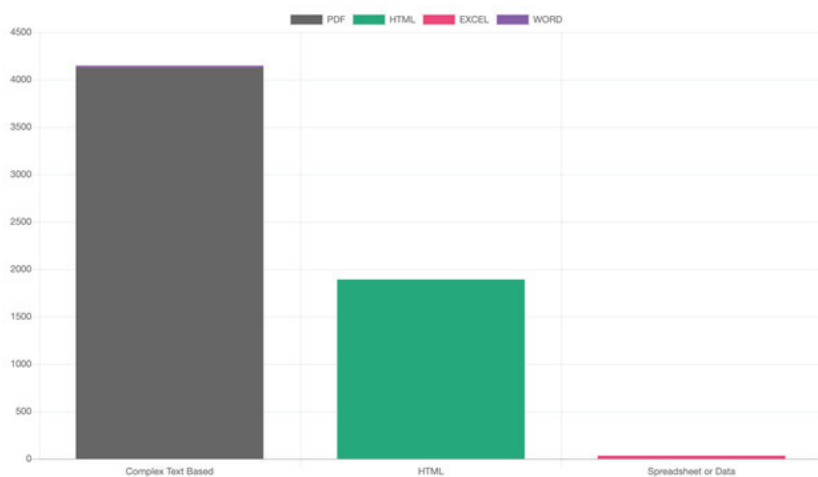
By graphing you have immediately increased the value of your auditing efforts. Fortunately, doing quick graphing is easy (assuming you have regular information about your content, such as what would be generated from a spider tool) to start.

Spider, Spreadsheet, Graphs, and ETL → A dynamic audit that facilitates exploration

A dynamic audit should include some or all of the following characteristics:

- Adding new metadata as you decide you need it (in this case an ETL tool can be helpful in merging the data)
- Defining and running rules to make decisions about content (an ETL tool can help here, as can a spreadsheet — this is perhaps the most difficult to implement)
- Scraping information off pages
- Pulling information out of URLs (such as the “folders” in the URL)
- Ability to graph and sample the content by clicking on the graph

The ETL tool can help you, in a more repeatable manner, add information about your content. For instance, I find it useful to break down content into “Complex Text Based”, “HTML”, and “Spreadsheet or Data” (rather than very detailed things like MIME-Type) like in this graph:



Example from Content Chimera

For instance, I have developed this type of audit using ScreamingFrog, Excel, MySQL, Tableau, and EasyMorph.

Although this type of audit is significantly more effective than the other types above, it is also significantly more difficult to implement and maintain (this is why I have developed a new web-based tool that does this rather than cobbling together other tools: Content Chimera <https://chimera.davidhobbsconsulting.com>).

Search as a service

Miles Kehoe

Introduction

Traditionally search has been an on-premise application but over the last few years the number of vendors offering a choice between an on-premise and a cloud application has increased substantially. One of the major announcements in 2018 was the launch by Google of its Cloud Search service as a replacement for its on-premise search appliance offering. Microsoft already offers Azure as a cloud search application and one of the other major players is Amazon Web Services. In addition, many search vendors offer a cloud version of the application and of course cloud content services vendors such as Box embed search into their service offering.

A brief history of hosted search

Hosted site search is not new; at least two companies, SearchButton.com and Atomz, introduced the concept in the 'Dot-Com' days of the late 1990s. Both companies offered similar services based on technologies including Verity and proprietary technology. Back then, the downfall was the 'give it away' mentality popular at the time; so, when things got tough, interest in hosted search faded.

Jump forward to the present, and there are more than a dozen companies offering professional quality search capability all at pretty reasonable rates. Early players of this generation, like Algolia and Swiftype, have excellent products, and the competition has followed suit. While some of these are free search services, a majority of the current vendors have learned the lessons of the past and license use of the technology for generally reasonable monthly fees.

Benefits of hosted search

Why rent an enterprise service when you can license it outright?

Like many other enterprise software tools, installing, configuring, and maintaining search in-house can be a challenge to your IT staff. First, not many enterprise staff have deep experience with search. And the cost of the equipment needed to drive the search application and configure it for failover and for high availability can be an expensive proposition.

When search is delivered as a service, the total cost of ownership drops. There are no servers to provision; load balancing and failover are built in; and the hosting company is responsible for monitoring the servers and insuring 7x24x365 reliability. The application is deployed quickly, in some cases in minutes, and at a quite reasonable price.

When you 'rent' a complex application like search, the IT overhead drops. However, this does not imply that you don't need any staff at all. As with any enterprise application, site search is not 'fire and forget'. Search, perhaps more than any other enterprise application, provides insight into users' intent as they access your site. A search team can use that knowledge to ensure the right content is available through search; to identify missing content; to review query activity; to manage 'best bets' and promotions; and to create reports for key stakeholders and product-line owners.

For a relatively low cost, you enjoy large infrastructure features at a low price. And while that's nice, low cost of ownership isn't the only benefit. Some other benefits that come along, include rapid deployment, low maintenance, good site reporting and operational support for server management.

Some of the available commercial solutions are appropriate for site search; others have sufficient capability to work properly for both site search and as a cloud solution.

Cloud search

Cloud search technology offers more capability, and is powerful enough to handle large enterprise, web, and eCommerce requirements. Cloud search requires some knowledge of HTML, but generally they are implemented via an API; and familiarity with at least web programming technologies, if not actual coding, is required to get everything working properly.

In return for the complexity, the technology is remarkably scalable, very fast, and suitable for even enterprise level requirements.

Cloud vendors

A number of enterprise and site search vendors now offer a cloud version of their product in addition to their conventional enterprise/site search offerings. Others, like Amazon, offer a cloud-only product.

Some companies offer a cloud option based on their enterprise product – Coveo and Lucidworks come to mind. This means that organisations familiar with conventional enterprise search will feel right at home; the only real difference is that the software resides in the cloud. The content may also be cloud based, but more likely the content remains within the corporate firewall and the cloud-based software and the search index resides in the vendor cloud instance. The type of content that is supported is generally the same as the formats supported in the on-prem version; and a team familiar with enterprise search will feel right at home; the only real differences are that the software and the index is maintained remotely; and queries and results may be encrypted in transit.

However, some of the cloud providers have architected their cloud product differently, and it makes sense to research the differences. For example, Amazon's AWS Cloud Search requires that content be pushed to the cloud instance, with supported formats being JSON and XML only. AWS also assumes that any fields – title, author and the like – are transmitted in the same formats. AWS Cloud Search does not have a traditional 'crawler' or 'spider'.

Most cloud-based platforms, including Google, Coveo, Lucidworks and others, offer a traditional crawler with support for common enterprise formats including Microsoft Office formats, text, PDF and others. However, because the technology is changing so quickly, be certain the cloud solution you go with meets your content requirements. In fact, I'd suggest a proof of concept on any search technology, regardless of whether you are using an on-prem or cloud-based platform.

Cloud search use cases

As mentioned, Cloud Search is often based on the same core technology that solves site search requirements. Unlike Site Search, Cloud Search generally fits well where document and index security are critical. This means the Cloud Search is more appropriate for use cases where document security is required. Some of these include:

Intranets

Almost by definition, intranet sites are intended for internal users; and not all content is intended for every employee. By indexing content linked with appropriate security credentials, intranet search can limit access to only those documents a searcher has rights to view.

Sites with sensitive content

Even within an intranet, some content is more restricted. This is often the case when HR sites need to be search enabled; but the results presented may vary based on the security credentials of the person doing the search.

Customer support sites

Support sites that may include customer names and the problems reported, require a higher level of security. It's possible a client problem report could provide sufficient information for an unauthorised person to make intelligent guesses regarding otherwise confidential information. When support tickets are protected by user, the risk of a data breach is much lower.

Search team support

Another common business case for cloud applications is the fact that they can reduce staff costs because all the administration is being carried out automatically by the cloud service. All the evidence shows that even quite substantial enterprise search implementations can be run with 2-3 staff with a good background in information retrieval. Once the initial index has been run, enterprise search applications are usually very robust and need little in the way of maintenance and monitoring as far as system performance is concerned. If the service does fall over then no customers are going to notice, and employees have many other ways to track down information if the search application is not available.

Where on-premise staff are essential is in assessing the retrieval performance of the application, and this requires them to be working very closely with line-of-business managers and employees. This role is not something that a cloud service provider will support. Certainly, the cloud application will generate reports, but they are just the initial basis for consideration of how to improve relevance and ranking.

Risks

By definition, both site search and cloud search expose content outside of relatively safe intranet firewalls. As such, there is always a risk that content intended for specific audiences may be vulnerable to access or even modification by unauthorised audiences. The only incident we've seen to date involving search was a recent report that a Facebook attack was able to retrieve users' queries. On a public site like Facebook, knowing what users searched for could be embarrassing. In the enterprise, simply knowing what queries were submitted could expose confidential information – secret project code names, employee names and email addresses, and even more sensitive personnel information.

At a high level, virtually all search platforms work their magic by accessing two different files: one contains a list of all words in all indexed documents; and another that stores metadata about every indexed document such as Title, Author, and sometimes even file permissions. Access to these files allows a motivated intruder to reconstruct virtually every document in your search index.

These risks apply not only to your internal enterprise search, but also to the search platforms behind the site search and cloud providers. Generally, site search as we have defined it here does not contain much, if any confidential data; but with cloud search beginning to offer powerful intranet search, you need to be sure your content is safe.

Search team skills

Martin White and Agnes Molnar

Achieving high levels of search satisfaction is not just a technology challenge but a people challenge. The critical success factor is this. User satisfaction with search performance is a function of the number of people in the search support team. An outcome of research published by [Paul Cleverley and Simon Burnett](#) in 2018 is that the underlying causes of low search satisfaction fall into three categories, which are Technology, Information and Literacy.

Technology

Two aspects of the technical implementation always give rise to user annoyance. The first is the reliability of the search tool and the second is the ranking of results. An element of the perceived reliability is the consistency of response. A search undertaken on Monday and then repeated on Friday as a check may give rise to a quite different set of results. This could be an outcome of the index update cycle or a new repository or application being added. The list is almost endless.

Ranking is always going to be a challenge in search because every user has their own view on the relevance of a result. When it comes to enterprise search the concept of relevance and the value of relevance tuning needs to take into account a range of related issues. Enterprise search users balance three criteria when assessing search results. The first is the extent to which the document is nominally relevant to their query. We use the term 'nominally' because this is the extent to which the retrieval model has determined the relevance based (for example) on the BM25F algorithm's assessment of the text content of the document. The second is perceived information quality, which we cover below.

The third criteria is usability, in the context of (for example) whether the format of the document is one that is relevant to a need or is in a language that the user can understand. A PowerPoint file may well be highlighted as a highly relevant document (especially in a SharePoint search!) but the user is faced with being unsure of the back story to the presentation and whether (for example) the presentation was heavily criticised when presented to a team meeting.

These issues are best met by relevance engineers working closely with the IT team. There is a requirement to understand business requirements and the technical capabilities of the search application(s) and so the person in this role is well positioned to take on the role of being the Search Manager.

Information

This will be a very personal view, based on a knowledge of the organisation and the people within it. Trust is an element of this quality assessment and a user is very likely to select a document that has been written by someone they know and trust than a document from an unknown author, or an author who (according to the people search functionality) is no longer an employee. Currency is also important. This is where the Last Modified date can be very misleading – it is quite possible for a document to have been written in 2015 and then very recently edited because the department affiliation is no longer correct.

Very few organisations have standards, or even guidelines, for information quality. There are many aspects of information quality. There is of course no realistic hope of going back through perhaps millions of documents and removing all those that do not meet quality standards. It is not just a question of resources but also the lack of definition of quality standards, or even quality guidelines.

In the case of web sites, content owners want their information to be found and used. Web teams and intranet teams often develop guidelines on how web information should be presented but most enterprise content is in Microsoft Office files. Very few employees think about the importance of a document they are writing and how others may find and use it. PowerPoint files are usually good examples where the title is often vague, and the content of each slide assumes a certain level of knowledge on the part of the reader. Setting guidelines for titles can make a significant difference and here consistency is important as search applications tend to weight the words in the title.

Finding people with the skills in metadata and taxonomy management can be very challenging as these are very specialised skills. Often these skills are needed on a periodic basis rather than on a day-by-day basis, which is why creating a virtual search centre of people with these skills can be a very beneficial investment for an organisation of almost any size.

The role of the Search Information Specialist is to develop guidelines for information quality that over time will result in a greater degree of trust in the information that users find. An important aspect of this work is to define some search personas so that the search solution can be customised to a specific group of users. In a pharmaceutical company, research scientists are going to use search applications in a very different way to pharmacovigilance teams.

Literacy

Search cannot be intuitive to every user. This is why some degree of customisation can be helpful. However, users do need to be supported in how to develop queries which balance recall and precision. All too often the approach taken is to assume that users will just use a single word query and then use filters and facets to drill down to relevant content. This is time-consuming and often frustrating. This is where the 2d Search tool developed by Tony Russell-Rose can be very useful in enabling users to test out different query options.

The support requirements are significantly greater when enterprise search is rolled out globally. There is likely to be a need for a Search Information Specialist for each major content language to identify any issues arising from poor stemming performance and inappropriate metadata tagging. This may not be a full-time position but certainly the expertise needs to be available to the search team. For similar reasons a good case can be made for an analytics specialist for each business area in a highly diversified global corporation.

Ideally there should be a Search User Support Manager in each major country, or at least each region (Europe, Asia/Pacific, North America) and language issues must be borne in mind. Although people may well speak several languages in business situations, they will prefer to search in the language in which they have the best command, so Spanish search and support in South America is very important, as of course is Portuguese for Brazil.

Managing the development process

Larger companies are now moving towards a two-stream programme for search development.

Stream 1 (Operations) supports the ongoing operational effectiveness of search. Multiple search evaluation techniques are used to identify potential search issues, ensuring that through multiple search evaluation techniques an early warning is gained of potential search issues. In addition, Stream 1 also assesses and prioritises user and business requirements.

Stream 2 (Development) is a development stream, where requirements identified in Stream 1 but which are beyond the skills and time of the search team to address are put out to specialist consultants and contractors.

The Global Search Manager acts as the coordinator of the two Streams and reports to the budget holder for search and to a Search Governance Committee.

Search team or search technology – which comes first?

The evidence suggests that almost all search applications will work substantially better when there is a search team with the skills and experience to enhance search performance to meet business and user requirements. These skills are so important that any company considering upgrading or replacing their current search applications should not do so unless there is at minimum a Search Development Manager to define the technical and user requirements and manage the selection and implementation process.

Training enterprise search managers

The demand for experienced enterprise search managers is significantly in excess of their availability. As a result they can command salaries (in the UK) of the order of £100k, placing them towards the very top of operational IT staff. There is virtually no training available for search managers, and in academic institutions the emphasis is inevitably on information retrieval. In December 2019 Professor Marteen de Rijk and Associate Professor Ilya Markov (University of Amsterdam) published an Opinion Paper in [SIGIR Forum](#) that set out a core discipline scope for information retrieval courses. It is disappointing that this paper made no reference to the two books on enterprise search by [Martin White](#) and by [Udo Kruschwitz](#) and [Charlie Hull](#), both of which were written as course textbooks for students as well as practical manuals for enterprise search managers. Currently the only training available for search managers on a commercial basis comes from [Search Explained](#).

Achieving enterprise search satisfaction

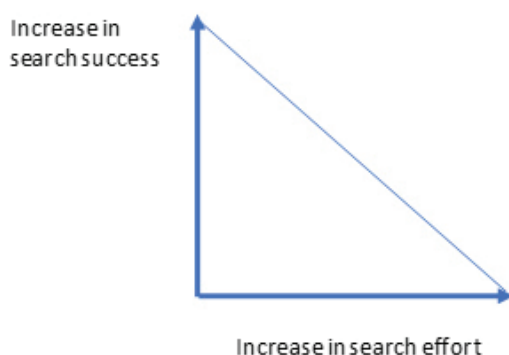
Martin White and Agnes Molnar

Enterprise search can trace its origins back to 1951 and enterprise search applications have been on the market since the mid-1980s, so by any measure this is a mature technology and a mature industry. It is therefore surprising and concerning that surveys have indicated that employees are experiencing considerable difficulties in getting the best from search applications. When conducting interviews for an intranet project, managers very quickly turn the conversation towards poor quality search and in twenty years of listening to conference presentations about enterprise search enhancement we have never yet heard a search manager describe dramatic growth in the number of queries and in search satisfaction.

An immense amount of research has been conducted on web search. The number of research papers is in the thousands. For comparison there are perhaps only a dozen or so research papers on enterprise search. A major reason for this is that it is very difficult to undertake research inside of organisations. The vacuum has been filled with a range of anecdotal pronouncements about what it takes to deliver high-quality enterprise search, with a strong focus on technology and on 'time saved'.

It is quite noticeable that in presentations at conferences (or even in the corridors afterwards) speakers are very reluctant to disclose the volume of queries they are achieving. This might indicate that query levels are much lower than they anticipated. Quoting increases in these levels following the adoption of a new search application would surely be in the interests of both the search vendor and the corporate search team. Given the maturity of enterprise search it is surprising that surveys from [AIIM](#), [Findwise](#) and [Net-JMC](#) over the last five years all indicate that users are finding it very difficult to locate the information they need.

There is now increasing interest in using 'search satisfaction' as a high-level metric of overall search performance from a user perspective. Search satisfaction is a function of search success against search effort.



Users are prepared to put effort into a search but at the same time are conscious of the amount of time and effort (and skill) that could be needed to achieve the search objective. There is a point at which the user makes a trade-off between effort and success, usually at a point where enough information has been gained to reduce the potential business risk of a decision to an acceptable level. Rarely is it necessary to find all relevant documents (100% recall), especially as 'relevance' itself is a subjective metric.

Although many people have offered suggestions for why this is the case it was not until the publication of a [research paper](#) by Paul Cleverley and Simon Burnett in June 2018 that the range and categorisation of search dissatisfaction became clear for the first time. The methodology is what is usually referred to as a longitudinal mixed methods approach. First, feedback was obtained from the search user-interface to gauge satisfaction with the search outcomes. Second, interviews were carried out with members of the thirteen internal and contract staff supporting the search application. The two data sets were then triangulated to highlight areas of agreement (all but two), dissonance (none) and silence (two). The study was longitudinal, with the same group of users being monitored over a period of two years. The interviews were coded so that a clear differentiation could be created between satisfaction and dissatisfaction.

In the paper the three factors identified that predominately influenced satisfaction were technology, information quality and information literacy and task utility. The technology factors include search tool reliability, search ranking and query syntax handling. In total these factors were the largest single group (38%), and that could be used as a justification for investing further in search technology. However together information factors (36%) and literacy factors (26%) accounted for 62% of the reasons for dissatisfaction and this indicates that technology investment on its own is not going to make a significant difference to search satisfaction.

Moving on to search-level metrics, the search application was used by around 70,000 staff each month and generated over 450,000 search queries. The average query length was 1.89 words and the top 30 most frequent searches fell from 14% of all search queries at the start of the project to just 8% at the end of the project two years later when of course users had gained substantially more experience with the application. This confirms anecdotal evidence that the tail of low frequency queries is very long in the enterprise environment. In our view this has significant implications for 'cognitive search' because there will be such low levels of use data from the majority of the queries that it will be hard predict optimal results. The percentage of results with 'no results' decreased from 0.4% to 0.3% over the same period. These metrics are the baseline that search managers have been seeking for years without success.

Many different approaches have been tried out to understand the context of enterprise search, including self-completion surveys or interviews at the end of a morning or afternoon. The problem with these approaches is that search users are relying on their memory of how long they spent seeking and/or searching. It is very difficult to obtain reliable and consistent results.

Now computational ethnography is being used to track how employees are seeking information, and what the role is for search in the process of seeking. This differentiation between seeking and searching is important. Employees have a range of ways of seeking information and is important to understand why they decide to use a search application. For too long the focus has just been on the technology of search even though the impact of context on search has been recognised and researched for over two decades.

In 1999 [Professor Tom Wilson](#), working at the Information School in Sheffield, developed a very useful schematic for the positioning of information behaviour, seeking and searching.



This positions information search behaviour, the process of using a search application as just one element of information-seeking behaviour, and that in turn reflects organisational information behaviours. The definitive analysis of information behaviours in organisations is 'The Inquiring Organisation' by Professor Chun Wei Choo, published in 2016. The sub-title of the book is 'How Organisations Acquire Knowledge and Seek Information'.

Using a search application is just one way to seek out information. When users realise they don't seem to have all the information they need to reach a decision, they may seek to fill this information gap by (as examples)

- Reading through documents we have stored on our personal or team files
- Using an enterprise application (HR, ERP, e-Learning etc)
- Sending an email to one or more people we know
- Taking to a colleague or an acknowledged expert
- Posting a request on a social media channel
- Browsing through an intranet
- Checking through a department or team wiki
- Asking for assistance at the next team meeting
- Searching on the web
- Searching on a specific application
- Searching across multiple applications

The act of searching must be put into this wider context so that we not only know how employees search but why they choose search as their option and what they then do with the information they find.

Over the last few years academic research teams, primarily in Finland and the UK, have started to examine in detail how employees go about their daily work within the context of seeking and searching for information.

Professor Järvelin has published three very important [papers](#) co-authored with one of his PhD students Miamaria Saastamoinen. The approach they have taken marks a breakthrough in enterprise search research in that they logged the use made of digital resources in the work place, rather than relying solely on diaries or self-completed surveys. The logging software provided 'dwell time' information, defined as the time in seconds that the participant kept the resource open as an active window during the work task. The log files were supplemented by a limited amount of workplace shadowing to provide a context to the way in which the digital resources were being used. The study participants were also asked to record information about the complexity of the tasks they were undertaking, and these tasks were also broadly categorised into communication, support, editing and intellectual tasks.

To quote from Professor Järvelin

"To design information search systems to properly serve WT [Work Task] requirements, it is necessary to study how the WTs as actions are connected to searching. Therefore, information (retrieval) systems development and evaluation should not take place in isolation but take the work context into account and find out for what purposes and how the systems are used. Failing to do this may result in developing suboptimal systems for expected but biased search needs."

This statement goes to the core of the likely causes of search dissatisfaction. Without understanding the relationship of work tasks and related search tasks in implementing enterprise search applications there is a substantial risk that the application will be fit to the specification (which, as an IT application focuses on functionality and technical performance) and yet not be fit for the purpose of providing an effective and satisfactory application for users.

In October 2018 The Search Network released a report [Achieving Enterprise Search Satisfaction](#). This report considered in detail how best to achieve search satisfaction. The graphic on p48 presents a summary of the report.

Application Capability

Inadequate technical capability is the largest single cause of search dissatisfaction. Fundamental shortcomings in the search application cannot be overcome by any amount of training and support. You are building on sand!

1. Seeking strategies

Employees have a wide range of options to find the information they need, such as sending an email to a colleague or raising the topic at a team meeting. Search should be positioned within this spectrum to meet specific requirements and not as the only solution

2. Information quality

All employees should take personal responsibility for information quality, making sure (for example) that titles are informative and that appropriate metadata is applied. This contribution has to be recognised and supported by their managers. Information is a corporate asset.

3. Meaningful metrics

Search logs are valuable in improving relevance ranking. If search satisfaction is low then the number of users will also be low. As a result the log data may not be representative. Log data will not reveal what stopping strategies are being used. Both qualitative and quantitative metrics need to be employed

4. Search personas

Search personas should reflect why employees are searching, and distinguish between searching for 'documents' and searching for knowledge and expertise. It is all about the business and personal context around the query. The analysis of search metrics will be enhanced by taking a persona perspective

5. Search team

Search team members have two roles. The first role is to provide support and training to employees, and gain feedback from interviews, surveys and log analysis. The second role is to enhance relevance ranking and the overall technical performance of the application based on search metric analysis and user feedback.

6. Customised solutions

It is very unlikely that a single user interface will be suitable for all users. The user experience cannot be 'intuitive' for everyone. A programme of user testing is essential as it will identify personas which would benefit from a customized solution

7. High visibility

The search team has to be highly visible and be proactive in providing guidance on how to get the best out of a search application. It has to have the capacity to resolve quickly instances of poor search satisfaction

8. Train and support

All enterprise applications require users to be trained and supported. Enterprise search is no exception. Few users have the skills needed to construct alternate queries when an initial attempt fails to deliver relevant content

Appendix A Enterprise search software

This list is included only to provide a starting point in creating a shortlist for an enterprise search project. There is no implied endorsement by members of The Search Network including many specialist software vendors, can be found at

<http://www.enterprisesearchbook.com/vendors/vendors-directory/>

A list of some search integration companies can be found at

<http://www.enterprisesearchbook.com/vendors/implementers/>

Company	HQ	Category	URL
Algolia	USA	SaaS	https://www.algolia.com
Amazon	USA	SaaS	https://aws.amazon.com/cloudsearch/
Attivio	USA	Commercial	http://www.attivio.com
Autonomy	UK	Commercial	https://www.microfocus.com/en-us/products/information-data-analytics-idol/resources
BAInsight	USA	Commercial	https://www.bainsight.com
Cludo	Denmark	Commercial	www.cludo.com
Coveo	USA	Commercial	http://www.coveo.com
dTSearch	USA	Commercial	http://www.dtsearch.com/
Elastic	Netherlands	Open Source	https://www.elastic.co/products/elasticsearch
Exalead	France	Commercial	https://www.3ds.com/products-services/exalead/products/
Findwise	Sweden	Open Source	http://www.findwise.com
Flax	UK	Open Source	http://www.flax.co.uk
Funnelback	N/A	Commercial	http://www.funnelback.com
Google	USA	Commercial	https://cloud.google.com/products/search/
Hyland	USA	Commercial	http://www.hyland.com/en/products/enterprise-search
IBM Watson	USA	Commercial	https://www.ibm.com/watson
IntraFind	Germany	Open Source*	https://www.intrafind.de/index_en
Lucene	N/A	Open Source	https://lucene.apache.org/
Lucidworks	USA	Open Source*	http://www.lucidworks.com
M-Files	Finland	Commercial	https://www.m-files.com
Mindbreeze	Austria	Appliance	http://www.mindbreeze.com
Microsoft SharePoint	USA	Commercial^	https://docs.microsoft.com/en-us/sharepoint/dev/general-development/search-in-sharepoint
Microsoft Azure	USA	SaaS	https://azure.microsoft.com/en-us/services/search/
Open Source Connections	USA	Open Source	http://opensourceconnections.com/
OpenText	Canada	Commercial	https://www.opentext.com/what-we-do/products/discovery
Oracle Secure Search	USA	Commercial	http://www.oracle.com/technetwork/search/oses/overview/index.html
RAVN	UK	Commercial	https://imanage.com/product/ravn/
Searchblox	USA	SaaS	https://www.searchblox.com/
Searchify	USA	SaaS	http://www.searchify.com/
Sinequa	France	Commercial	http://www.sinequa.com
Squirro	Switzerland	Commercial	https://squirro.com/

Company	HQ	Category	URL
Solr	N/A	Open Source	http://lucene.apache.org/solr/
Swifttype	USA	SaaS	https://swifttype.com/
Vespa	USA	Open Source	http://vespa.ai/
Yippy	USA	Commercial	https://yippy.com/
Voyager	USA	Commercial	http://www.voyagersearch.com

Notes

Findwise, IntraFind and Lucidworks are based around Lucene, Solr and in the case of IntraFind Elasticsearch. However these companies also integrate modules which are provided on a commercial basis, and so in effect are a hybrid of open source and commercial products. IBM, Microsoft and Oracle do not offer stand-alone search applications.

Search resources: books and blogs

The books listed below represent a core library which should be on the bookshelf of any manager with enterprise search responsibilities.

Designing the Search Experience

Tony Russell-Rose and Tyler Tate, 2012. Morgan Kaufmann ([Book website](#)) ([Review](#))

This book takes a deeper look into information seeking models, using them to consider how best to design user interfaces.

Enterprise Search

Martin White, 2nd Edition 2015. O'Reilly Media ([Book website](#))

My objective was to write a book for search managers without a technical background that supported the entire process from building a business case through to evaluating performance.

The Inquiring Organisation

Chun Wei Choo, 2015. Oxford University Press ([Review](#))

The importance of this book is that it provides a context for search within an overall integration of the value of information and knowledge to the organisation.

Interactions with Search Systems

Ryen W. White, 2016. Cambridge University Press ([Review](#))

Although the focus of this book is on web search, the principles also apply to e-commerce and enterprise search.

Introduction to Information Behaviour

Nigel Ford, 2015. Facet Publishing ([Review](#))

Information seeking models are a special case of information behaviours. They form the basis of use cases for search, and the design of user interfaces.

Looking for Information

Donald O. Case and Lisa M. Given, 4th Edition 2016. Emerald Publishing ([Book website](#))

A survey of research on information seeking, needs and behaviour, which places search into the wider context of why people seek information and how they interact with search systems.

Multilingual Information Retrieval

Carol Peters, Martin Braschler and Paul Clough, 2012. Springer ([Book website](#))

A good introduction to the basic principles of multilingual and cross-lingual search.

Relevant Search

Doug Turnbull and John Berryman, 2015. Manning Publications ([Book website](#)) ([Review](#))

The objective of all search applications is to deliver the most relevant results as early as possible in the list of results. Although based around the management of Lucene and Solr this book is applicable to any search application.

Search Analytics for Your Site

Louis Rosenfeld, 2011. Rosenfeld Media ([Review](#))

This introduction to search analytics is primarily about websites and intranets but the principles apply to enterprise search.

Searching the Enterprise

Udo Kruschwitz and Charlie Hull, 2017. Now Publishers ([Review](#))

The authors provide an important bridge between information retrieval research and the practical implementation of search applications.

Text Data Management and Analysis

ChengXiang Zhai and Sean Massung, 2016. ACM/Morgan & Claypool ([Review](#))

A very comprehensive handbook on the technology of information retrieval and content analytics based on a highly regarded MOOC.

[Morgan & Claypool](#) and [Now Publishers](#) both offer a wide range of books on specialist aspects of information retrieval and search, though with an academic rather than a practitioner focus.

This is a list of blogs whose authors comment on aspects of search technology and implementation on a reasonably frequent basis.

[All About Search](#) Ronald Baan
[Beyond Search](#) Stephen Arnold
[Complex Discovery](#) Rob Robinson
[Concept Searching](#) Corporate blog
[Coveo Insights](#) Corporate Blog
[Daniel Tunkelang](#)
[Data Dexterity](#) Corporate blog for Attivio
[Do More With Search](#) BA Insight corporate blog
[Elastic](#) Corporate blog
[Enterprise Search](#) Miles Kehoe
[Flax](#) Charlie Hull
[Funnelback](#) Corporate blog
[Information Interaction](#) Tony Russell-Rose
[Intranet Focus](#) Martin White
[LucidWorks](#) Corporate blog
[Matt McDermott](#)
[Opensource Connections](#) Corporate blog
[Searchblox](#) Corporate blog
[Search and Big Data Insights](#) Paul Nelson, Search Technologies
[Search Explained](#) Agnes Molnar
[Sease](#) Corporate blog
[Sinequa](#) Corporate blog
[Synaptica](#) Corporate blog
[Systems Thinking](#) Paul Cleverley
[Tech and Me](#) Mikael Svenson

In addition, the [Special Interest Group on Information Retrieval](#) of the British Computer Society and the [Special Interest Group on Information Retrieval](#) of the Association for Computing Machinery publish newsletters.

Glossary

Absolute boosting

Ensuring that a specified document always appears at the same point in a results set, or always appears on the first page of results.

Access control list (ACL)

Defines permissions to access a specific repository, a set of documents, or a section of a document.

Advanced search

The provision of a search user interface which prompts the user to enter additional terms to assist in ranking results, often using Boolean operators.

Apache

The Apache Foundation provides support for a wide range of open source applications, including Lucene and Solr.

Appliance

A search application pre-installed on a server ready for insertion into a standard server rack.

Auto-categorisation

An automated process for creating a classification system (or taxonomy) from a collection of nominally related documents.

Auto-classification

An automated process for assigning metadata or index values to documents, usually in conjunction with an existing taxonomy.

Average response time

An average of the time taken for the search engine to respond to a query, or the average end-to-end time of a query.

Best bets

Results that are selected to appear at the top of a list of results that provide a context for other documents generated and ranked by the search application.

BM25

A ranking function developed in the 1990s but still widely used. It has its origins in the tf.idf ranking function.

Boolean Operators

A widely used approach to create search queries; examples include AND, OR, and NOT—for example, information AND management.

Boolean search

A search query using Boolean Operators.

Boosting

Changing search ranking parameters to ensure that certain documents or categories of documents appear in the results.

Categorisation

The placing of boundaries around objects that share similarities (e.g., taxonomy).

Clustering

A process employed to generate groupings of related words by identifying patterns in a document index.

Cognitive search

A description loosely applied by search vendors to applications using machine learning and AI techniques to determine the work context of the user and deliver personalised results.

Collection

A group of objects methodically sorted and placed into a category.

Computational linguistics

The use of computer-based statistical analysis of language to determine patterns and rules that aid semantic understanding.

Concept extraction

The process of determining concepts from text using linguistic analysis.

Connector

A software application that enables a search application to index content in another application.

Controlled vocabulary

An organised list of words, phrases, or some other set employed to identify and retrieve documents.

COTS

Commercial off-the-shelf software.

Crawler

A program used to index documents.

Cross-language search

A query in one language is translated into other indexed languages (often using a multi-lingual thesaurus) so that all documents relevant to the concept of the query are returned no matter what language is used for the content.

Description

A brief summary, generated automatically, that is then included as a description of a document in the list of results. *See also Key sentence*

Document

A structured sequence of text information, but often used as a generic description of any content item in a search application.

Document processing

The deconstruction of a document into a form that can be tokenised and indexed.

Document repository

A site where source documents or other content objects are stored, generally a folder or folders. *See also Information source*

Early binding

A search conducted only across documents that a user has permission to access. *See also Late binding*

Entity extraction

The automatic detection of defined items in a document, such as dates, times, locations, names, and acronyms.

Exact match

Two or more words considered mutually inclusive in a search, often by enclosing them in quotation marks—for example, “United Nations”.

Facet

Presentation of topic categories on the search user interface to support the refinement of a search query.

Fallout

A quantity representing the percentage of irrelevant hits retrieved in a search.

Federated search

A search carried out across multiple repositories and/or applications.

Field query

A search that is limited to a specific field in a document (e.g., a title or date).

Filter

A function that sets specific criteria for search results.

Freshness

The time period between a document being crawled and the index being updated so that a user will be able to find the document.

Fuzzy search

A search allowing a degree of flexibility for generating hits (i.e., matches that are phonetically or typographically similar).

Golden set

A set of documents used to benchmark search performance that is representative of content that will be searched on a regular basis.

Guided search

A search in which the system prompts the user for information that will refine the search results.

Hit

A search result matching given criteria; sometimes used to denote the number of occurrences of a search term in a document.

Index

List containing data and/or metadata indicating the identity and location of a given file or document.

Index file

A file that stores data in a format capable of retrieval by a search engine.

Ingestion rate

The rate at which documents can be indexed, usually specified in Gb/sec.

Inverse document frequency (IDF)

A measure of the rarity of a given term in a file or document collection.

Inverted file

A list of the words contained within a set of documents, and which document each word is present in, so acting as a pointer to a document.

Inverted index

An index whose entries identify a given word and the documents in which it appears.

Iterative calculation

A calculation utilising a recursive and self-referential algorithm.

Key sentence

A brief statement that effectively summarises a document, often employed to annotate search results.

Keyword

A word used in a query to search for documents.

Keyword search

A search that compares an input word against an index and returns matching results.

Language detection

The indexing process identifies the language (or languages) of the content and assigns it to appropriate language specific indexes.

Late binding

Access permission checking carried out immediately before the presentation of the document to the user. *See also Early binding*

Lematisation

A process that identifies the root form of words contained within a given document based on grammatical analysis (e.g., run from running). *See also Stemming*

Lexical analysis

An analysis that reduces text to a set of discrete words, sentences, and paragraphs.

Linguistics

The study of the structure, use, and development of language.

Linguistic indexing

The classification of a set of words into grammatical classes, such as nouns or verbs.

Meta tag

An HTML command located within the header of a website that displays additional or referential data not present on the page itself.

Metadata

Data that provides information about other data (i.e., is data about data).

Morphologic analysis

The analysis of the structure of language.

Natural language processing

A process that identifies content by attempting to adhere to the rules of a given language.

Natural language query

A search input entered using conventional language (e.g., a sentence).

Parametric search

A search that adheres to predefined attributes present within a given data source.

Parsing

The process of analysing text to determine its semantic structure.

Pattern matching

A type of matching that recognises naturally occurring patterns (word usage, frequency of use, etc.) within a document.

Phrase extraction

The procurement of linguistic concepts, generally phrases, from a given document.

Precision

The quantification of the number of relevant documents returned in a given search.

Proximity searching

A search whose results are returned based on the proximity of given words (e.g., 'pressure' within four words of 'testing').

Query by example

A search in which a previously returned result is used to obtain similar results.

Query transformation

The process of analysing the semantic structure of a query prior to processing in order to improve search performance.

Ranking

A value assigned to a specific result returned for a query—the first item listed has a ranking of 1, the second has a ranking of 2, and so on.

Recall

A percentage representing the relationship between correct results generated by a query and the total number of correct results within an index.

Relevance

The value that a user places on a specific document or item of information. Both precision and recall are defined in terms of relevance.

Search results

The documents or data that are returned from a search.

Search terms

The terms used within a search field.

Semantic analysis

An analysis based upon grammatical or syntactical constraints that attempts to decipher information contained in a document.

Sentiment analysis

The use of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in documents.

Soundex search

A search in which users receive results that are phonetically similar to their query.

Spider

An automated process that provides documents to a data extraction or parsing engine. *See also Crawler*

Stemming

A process based on a set of heuristic rules that identifies the root form of words contained within a given document (e.g., run from running). *See also Lemmatisation*

Stop words

Words that are deemed to have no value in an index. *See also Word exclusion*

Structured data

Data that can be represented according to specific descriptive parameters—for example, rows and columns in a relational database, or hierarchical nodes in an XML document or fragment.

Summarisation

An automated process for producing a short summary of a document and presenting it in the list of results.

Synonym expansion

Automatically expanding a search by adding synonyms of the query terms derived from a thesaurus.

Syntactic analysis

An analysis capable of associating a word with its respective part of speech by determining its context in a given statement.

Taxonomy

In respect to search, the broad categorisation of objects (typically a tree structure of classifications for a given set of objects) in order to make them easier to retrieve and possibly sort.

Term frequency

A quantity representing how often a term appears in a document.

TF.IDF

The term frequency-inverse document frequency formulation gives a score that is proportional to the number of times a word appears in the document offset by the frequency of the word in the collection of documents. *See also BM25*

Thesaurus

A collection of words in a cross-reference system that refers to multiple taxonomies and provides a kind of meta-classification, thereby facilitating document retrieval.

Tokenising

The process of identifying the elements of a sentence, such as phrases, words, abbreviations, and symbols, prior to the creation of an index.

Truncation

Removal of a prefix or suffix.

Unstructured information

Information that is without document or data structure (i.e., cannot be effectively decomposed into constituent elements or chunks for atomic storage and management).

Vector space

A model that enables documents to be ranked for relevance against a query by comparing an algebraic expression of a set of documents with that of the query.

Weight

A value applied to a given area of a search system (e.g., term weighting, which represents its importance with respect to other factors).

Wildcard

A notation, generally an asterisk or question mark, that when used in a query, represents all possible characters (e.g., a search for boo* would return book, boom, boot, etc.).

Word exclusion

A list containing words that will not be indexed—this usually is comprised of words that are excessively common (e.g., a, an, the, etc.).