Search Insights 2021

The Search Network

March 2021

Contents

Introduction	1
The challenge of information quality, Agnes Molnar	5
The art and craft of search auto-suggest, Avi Rappoport	9
The enterprise search user experience, Martin White	15
Searching fast and slow, Tony Russell-Rose	20
Reinventing a neglected taxonomy, Helen Lippell	26
Ecommerce site search is broken: how to fix it with open source software, Charlie Hull & Eric Pugh	29
The data-driven search team, Max Irwin	32
Search resources: books and blogs	37
Enterprise search vendors	40
Glossary	42

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

Introduction

Successful search implementations are not just about choosing the best technology. 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.

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. There are now eleven members spanning the world from Singapore to San Francisco. We have known each other for at least a decade and share a common passion for search that delivers business value through providing employees with access to information and knowledge that enables them to make decisions that benefit the organisation and their personal career objectives. 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.

Search Insights 2021 is our fourth annual report. 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 world-wide. That is why there is no charge for this report, and it carries no sponsorship.

When we work together to agree the scope of our contributions we always take a forward-looking perspective, highlighting developments and methodologies that will make a difference in the year ahead. E-commerce has benefitted from changes in the way we shop for home and business and we expect to see wider adoption of AI and machine learning to improve the customer experience. There is also a shift in enterprise search towards a focus on the user experience that can take advantage of significant developments in AI, machine learning and natural language processing. Despite the claims for these technologies taxonomies still have an important role to play.

No matter how sophisticated the technology the performance of search applications has a critical dependence on information quality. On the web people want their web pages to be found and many organisations pay constant attention to search engine optimisation. Inside the organisation where documents are primarily written for colleagues there are no incentives to go the extra mile and make documents more findable. There should be policies for information quality but that is rarely the case.

Whether you are involved in managing enterprise, e-commerce or web search the benefits of having a search support team are both immediate and long term. So often it is the technology that gets the blame for poor search but a lack of investment in a search team and inconsistent content quality cannot be ameliorated by technology alone. These are just some of the topics that we have covered in this issue.

We were tempted to say something about the state of play in open-source search given the publicity around the change in license policies by Elastic but it is far too soon to take a view on the implications. Whatever the outcome open-source applications are now widely adopted despite (as the list at the end of the report shows) competition from around 70 commercial enterprise search applications.

All organisations are now having to plan for periods of uncertainty, opportunity and challenge as we gradually emerge from the pandemic. One of the outcomes of the pandemic has been a realisation of how important it is to find trustworthy information on which to make decisions.

Given the high profile that the Presidency of the USA has had in 2o2o it might be worth recalling that in opening the 1974 conference of the International Federation for Information and Documentation President Lyndon Johnson commented that the fire of progress is lit by inspiration, fuelled by information and sustained by hope and hard work. Search technology has a vital role in the process of information discovery, and we look forward to having the opportunity of working with you to help you and your colleagues achieve search excellence.

The Search Network

You can download previous editions of our Search Insights report here:
Search Insights 2018
Search Insights 2019
Search Insights 2020

The Search Network

Charlie Hull, OpenSource Connections **(USA & UK)**
Charlie co-founded search consultancy Flax before joining OpenSource Connections where he acts as a Managing Consultant and leads operations in the UK. 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 @Flaxsearch.

Max Irwin, OpenSource Connections **(USA)**
Max is a Managing Consultant at OpenSource Connections, which aims to empower organisations and search teams through consulting, strategy, and training. He has deep practical expertise in search relevance, customer experience, natural language processing, and growing engineering culture. Follow Max on Twitter @binarymax or connect with him on LinkedIn.

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 Electronic Arts, Pearson, the BBC, gov.uk, Financial Times, Time Out, 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.

Maish Nichani, PebbleRoad Pte Ltd **(Singapore)**
Maish Nichani is co-founder of PebbleRoad, a strategy, design and innovation practice based in Singapore. He is on a mission to help well-established organisations thrive in a digital world. PebbleRoad helps design digital products and services, and often these are search-driven experiences. Maish finds that many organisations are not aware of the benefits of enterprise search and are missing on a lot. He gives talks and demos at every opportunity but hopes to do more.

Eric Pugh, OpenSource Connections **(USA)**
Eric Pugh is the co-founder and CEO of OpenSource Connections. He has been involved in the open source world as a developer, committer and user for the past fifteen years. He is a member of the Apache Software Foundation and an active committer to Apache Solr. He co-authored the book Apache Solr Enterprise Search Server, now on its third edition. He also stewards Quepid, an open source platform for assessing and improving your search relevance.

Avi Rappoport, Search Tools Consulting **(USA)**
Avi Rappoport has been working on improving search since 1998, having previously earned a Master's in Library and Information Studies and worked in small software startups. She has advised companies on multi-source internal enterprise search, site search, informational search, and high-traffic ecommerce large product catalogue search. You can follow Avi on Twitter @searchtools_avi.

Tony Russell-Rose, 2Dsearch **(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 Royal Academy of Engineering Visiting Professor at Essex University and Senior Lecturer in Computer Science at Goldsmiths, University of London. He 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.

Cedric Ulmer, France Labs **(France)**
Cedric is the CEO cofounder of France Labs, a startup specialised in open source search engines and maker of Datafari, an open source enterprise search solution. He manages the company and handles the innovation and marketing aspects. In terms of ecosystem, he manages the open source business community at the largest association for IT entities in the French Riviera. He has been teaching entrepreneurship for four years at the Data Science European MsC of the EIT. Prior to that, he spent ten years at SAP in the research department. Cedric holds the French grande ecole diploma from Telecom SudParis, with the Eurecom certificate.

Martin White, Intranet Focus Ltd **(UK)**
Martin is an information scientist who started working with search technology in 1975. Over the last two decades he has worked on a wide range of enterprise search projects in North America and Europe. Based on his project experience he has written four books on enterprise search management. Since 2002 he has been a Visiting Professor at the Information School, University of Sheffield, where he lectures on information retrieval and information management. Follow Martin on Twitter @IntranetFocus.

The challenge of information quality

Agnes Molnar, Search Explained

In many cases organisations ask search consultants to "make search work". And in most cases, it's considered to be a single and straightforward IT project: the consultant does some 'magic' and search just starts working.

However, in most cases, it's not that simple.

Nobody has the 'search magic wand'. Usually, we have to do significant work on the content before doing anything with search. With this case study, let me show you one very typical example which will demonstrate the value and challenges of information quality.

Note: The project is on Microsoft 365 but could be on any other enterprise system.

Background

An international organisation with offices in almost 50 countries realised that the more content it migrates to and stores in Microsoft 365, the worse the findability of everything becomes – despite the promise of out-of-the-box Microsoft Search.

There are some common mistakes organisations make when they migrate and create their content without planning and governance. The company in this example is not an exception: they had the following issues and challenges before we started to work with them:

- inconsistent information architecture
- missing metadata
- inconsistent metadata
- lack of content lifecycle
- lack of content curation
- inconsistent use of languages and translations.

Inconsistent information architecture

When there's no plan and no guidance, nobody knows where and how to store the content. People do their best, but everyone has different backgrounds, experiences, and knowledge – therefore the way they store their content will be different.

Some might create top-level containers (site collections) for their team documents – others save all of the company archive into a single folder structure (library).

Some store their collaboration files in SharePoint – others share high importance corporate policies from their personal OneDrive. And everyone creates teams and channels in Microsoft Teams without knowing the implications.

And when none of these applications can satisfy the users, new applications come, including shadow-apps, because the promise of these is always better than the messy reality. And this spiral of adding more and more applications gets worse over time, with no real long-term benefits. Some examples are shown on the next page.

- The organisation standard is to use Microsoft 365, but the users are not educated. They create SharePoint sites for everything, but they don't use the available function-

alities – instead, they store the content in multi-level embedded folders, with no metadata. Same as storing the documents in a network drive.
- The organisation opens Microsoft Teams to enhance collaboration. With no guidance or governance, thousands of teams and channels are created. Every user becomes a member of dozens of teams, and eventually, the communication becomes too noisy, and users stop using Teams completely.
- The organisation stores documents in SharePoint Online, but sharing is disabled. Users start to use OneDrive, and eventually OneDrive becomes the primary source of truth. And since OneDrive is personal and is not meant for teamwork, the organisation will face major issues when the file owner leaves the company.
- The organisation stores documents in SharePoint Online, but external sharing is disabled. The users recognise that there must be a better option than sending documents as email attachments – so they start using Dropbox or Google Docs to share content with external partners. In many cases, these (sensitive) documents can be found on the public internet, too.

Missing metadata

Another challenge is when even the most advanced content management systems are being used as "smart" file shares. Users store their documents there, maybe organised into folders, but with no metadata at all. When they need something, they navigate to the content through the folder structure. However, if they don't know where the content they need is stored, they're lost.

At the same time, folder structures follow some logic: if you ask the users, they can tell you the first level is the client, the second level is the year, the third level is the project, etc. But using explicit metadata instead of cascaded folders is not something they're familiar with or understand.

Some examples include:
- Many use folders instead of metadata. Storing year, client, customer, project name, project ID, etc. as metadata can provide much better filtering, sorting, ordering, grouping, and search options; therefore, the overall findability of content and user satisfaction will improve.
- In many cases, the metadata can be found in the document (implicit) but not added to the document as explicit metadata. While full-text search works to some extent in this case, explicit metadata can improve filtering, sorting, ordering, grouping, etc. options.

Inconsistent metadata

If there's something worse than no metadata, it's inconsistent metadata. Below are a few examples of what we find when doing content inventory at our clients:

- There's no managed taxonomy, and users enter various synonyms for the same term. For example, "The Search Network", "Search Network", "Search Netw.", "TSN".
- Inconsistent use of languages and translations. If a user knows everything is in a common language (for example, English), it's a very clear approach. However, if the organisation uses multiple languages, there must be a language strategy. What should be translated? What types of content are available locally? Also, if there's a taxonomy, it has to be multi-lingual so that everyone can use it consistently and coherently.
- All metadata are free text, with no guidance or governance. Everyone enters whatever value they want to. In many cases, they are not even consistent with themselves.

- Date fields are used inconsistently. When there's a "date" field, users' understanding of what the "date" is about might be different. For example, during the scoping workshops, we often find that there's only one "date" field assigned to the documents. Users use it to enter the date of approval, uploading, valid from, valid to, effective from, effective to, etc. – just to name a few, from the same environment. Different date formats also add to the complexity: in many cases, it's not obvious whether 1/2/2021 means Jan 2 or Feb 1. It's not only confusing when a user is entering the date but also when filtering or searching for a specific date range.

Lack of content lifecycle

Knowing which document is the "real", the "official", the most recent, or the approved one is essential. However, if there is no content lifecycle in place, and users send various versions of documents back and forth by email, there is no way to know which one is the "right" one: multiple versions, with different and often conflicting content, is very common.

Lack of content curation

Related to content lifecycle, content curation often fails, too. Even when the information architecture, taxonomy, and metadata are all in place, users make mistakes. This is why content curation should be part of the lifecycle: a formal process to review and correct information structure and metadata as needed.

When content curation is missing, information siloes and multiple document versions are created, and the quality of metadata becomes messier and messier. After a while, the result is an information jungle where nothing can be found.

Inconsistent use of languages and translations

In a multi-lingual organisation, localised and translated content are an integral part of operations. Not every document has to be translated and localised, but for the ones that need to be localised, a consistent and appropriate translation is a must. In many cases, the localised pages and documents are not synchronised and updated when the original (mostly English) page changes, resulting in the same inconsistent behaviour as inconsistent information architecture or inconsistent metadata described above.

The (false) hope and promise of AI and auto-tagging

In the last few years, we've seen the rise of intelligent classification and auto-tagging solutions. While these might work in specific domains, applying them in any generic use case is as risky as the conditions described above. AI models have to be trained, maintained, and curated – and this might require fewer resources in total and also more planning and preparation in advance. A training set of content has to be identified, tested, an AI model has to be taught, evaluated, tested – all this in an iterative way, to improve and enhance the AI model until the quality of tags and terms applied gets good enough.

And this leads us to the question: how to measure information quality? How do we know if the content is trustworthy and of good quality? How do we know if it's not? How do we measure the quality of information quality? How do we measure the quality of auto-tagging, and how do we compare it to human tagging? What metrics should we apply? – these questions always have to be answered before the implementation starts, otherwise how do we know if/when we're successful?

The solution

As you can see, the challenge is quite complex. Moreover, it is unique to every organisation – there is no one-size-fits-all solution. The best you can do is to undertake a detailed content inventory, evaluate and analyse what information your organisation has, and classify it by the following dimensions:

- type of content
- metadata requirements
- content lifecycle requirements
- permission and accessibility requirements.

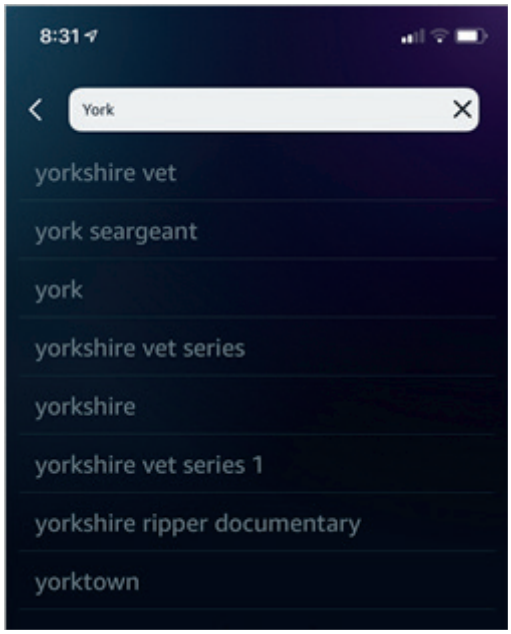
Once the inventory is done, define the priorities.

Set up SMART (Specific, Measurable, Achievable, Relevant, Time-bound) goals, and commit to short, mid-, and long-term processes to improve information quality. Don't rush this process, but take significant actions towards the desired goals. Measure often, and align your steps ahead as needed but always have a plan to follow. This leads you to your goals, step-by-step.

The art and craft of search auto-suggest

Avi Rappoport, Search Tools Consulting

Search auto-suggest - the drop-down menu suggestions displayed while typing into a search box - is the first and often the only search interaction that users encounter on a site or intranet. When done correctly, it is nearly invisible, offering both popular and rare search terms, and showing the vocabulary and scope of relevant searches. Web search engines such as Google have millions of queries and decades of search relevance statistics to generate these lists, but smaller sites can tune the suggestions and offer great value to users. While user activity can dynamically add new suggestion terms and re-rank depending on current value, this must be balanced with an ongoing evaluation and curation process to avoid suggesting inappropriate or misleading terms.



Amazon Prime auto-suggest for the word "york"

Note: auto-suggest functionality is different from interactive typing autocomplete, type ahead or word completion, because there are very likely to be multiple possible suggestions, and the result of a selection is not a simple substitution but an action to send a search or go to a specific content page.

Technically, auto-suggest is a simplified text search, with an index, retrieval, results ranking and presentation. However each of these is quite different from traditional document or product search.

Auto-suggest search index

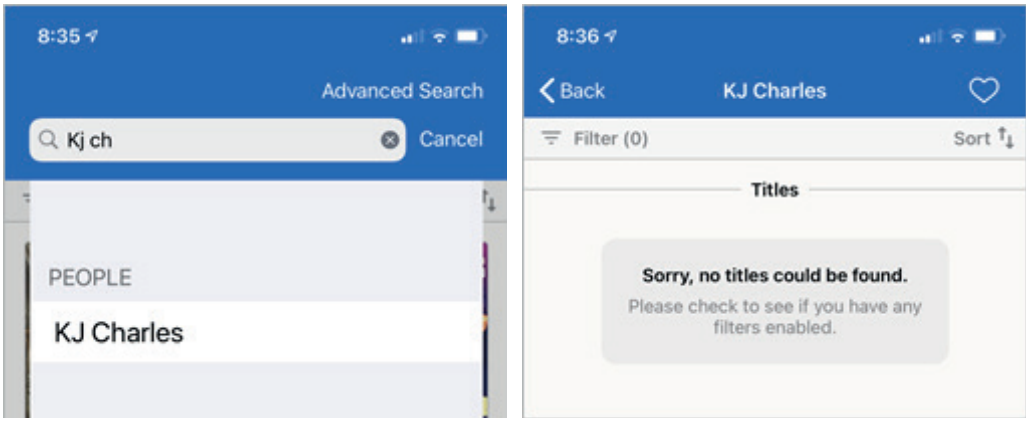
The internal representation of auto-suggest index terms comes in three main elements: Searchable text, Block lists, and Display text.

The **searchable text** is that matched when the user types and the front end sends the letter(s) to the auto-suggest backend. This is usually created from frequent successful user queries. In addition, **add lists** contain positive terms such as product names, brands, document titles and new vocabulary. Some of these, such as those dealing

with shipping or returns in a commerce site or a travel portal in an intranet, may redirect to specific URLs instead of sending a query to the search engine.

Block lists allow additional control, avoiding suggesting obsolete terms, unreleased products, out-of-scope queries and known misspellings. By managing these lists, domain experts and merchandisers can keep the suggestions current and avoid inappropriate suggestions.

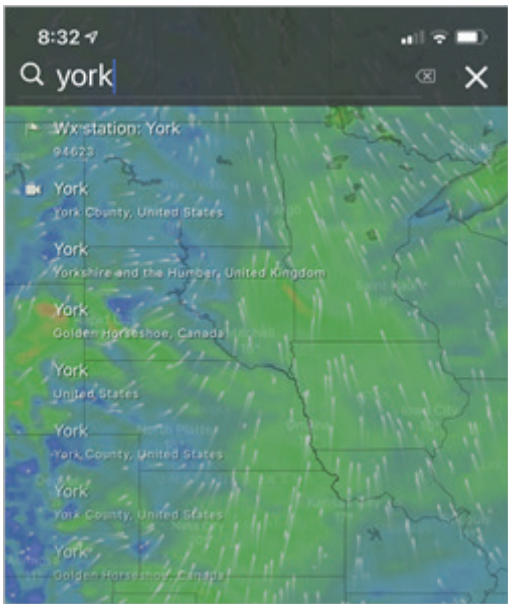
Note: it is extremely important not to mislead users by suggesting terms that have no matches in the search results. This generally happens when a user can only see part of a search index, for example a department or a specific section within a larger system, such as a particular city library holdings within a provider's ebook corpus. Adding a flag as part of the auto-suggest searchable query index, and sending this flag as part of the retrieval process avoids this disappointing user experience.



ebook app – the complete index has the author but the local library does not

This is even more important when terms related to sensitive topics such as unreleased products, policy changes, or employment should not be displayed unless the user has access permissions. Filtering on the user access permissions allows appropriate retrieval of this material, while avoiding the possibility of other users combining terms to identify topics that are not public, even if they cannot see the documents or records in search results.

The **display text** includes the correct capitalisation, singular/plural format, diacritics and punctuation, according to the requirements of the site. For example, a certain sporting event should be shown as the Super Bowl, even if a user types superbowl, and the French "thé" (tea) should not be shown as "the". This may also include variations on a term such as plum (fruit) and plum (colour), or a specific location in a geospatial search.



Windy.com auto-suggest list for the word "York", showing specific locations

Search URL and optional query parameters

The query text is the words from the suggestion to pass along to the full-text search engine. In most cases, this should send a phrase match setting for multi-word queries. This could also specify a field, permissions group or other filter, based on the display text.

For known items, such as brand names, categories or departments, the suggestion may simply redirect to the specific page.

Matching and retrieval

Suggestion query matching is unlike full-text search in that it generally performs a beginning (left) match on the searchable text, instead of requiring a full token (word) match. For example, a search for plu may match plum or pluot.

The most valuable suggestions are those with exact matches and matches at the beginning of the phrase. When there aren't any of those available, falling back to matching on the beginning of another word is still quite useful. For English-language suggestions, a partial match in the middle of a word is difficult to understand, however for many German words, it would be clear and helpful.

Beginning (left) matches
- **apple**
- **apple** sorbet
- **apple**-celery granita

Fallback token left matches
- green **apple** ice
- custard **apple** frozen yogurt

True substring match may be necessary in some cases
- Orangen**saft** (orange juice)

As described above, if there is a filter to what a user can see in search results, it's vital to include the same filter as a flag during suggestion retrieval.

Suggestions ranking

The order of suggestion results is not based on relevance, but on an internal value order. This can be precalculated using a set of algorithms based on search frequency and success metrics such as clicks and conversions, preferably calculated using ML features. There should also be boost factors for novelty, and decay factors based on the recent frequency and interactions. Some sites, such as food or decoration, may require higher decay factors to reduce the rank of outdated suggestions, such as santa claus in February.

In the frequent cases where there are ties on ranking values, subsorting the ranks alphabetically makes it easy for users to skim the displayed results.

Note that inappropriate suggestions, even innocuous ones, may be intriguing to users and cause them to select from the menu to see what happens. For example, there may be no ice cream called bandana, but if customers type the term often enough, it may rise to appear in the suggestion band. Once suggested, customers may want to know what it means, and select it, creating a problem feedback loop. To avoid this, the ranking algorithm should track the searches and subsequent conversions very closely, and alert domain experts about anomalies for possible addition to the searchable index or to be added to the block list.

Once the list is generated in order, it's a simple matter to retrieve the first ten (or so) left-matching suggestions. Even on web search and very large commerce sites, the number of choices generally limits itself quickly.

banana
banana cream
Bailey's
Bavarian cream
Bananas Foster
banana strawberry
banana nut bread
apples and **ba**nanas

As explained earlier, if there aren't enough matches on the left, re-search allowing left matches on other tokens, and rank those after the left matches.

The returned content should be the "display text" versions of the terms, with optional query parameters to send to the full-text search engine. The search engine may redirect these to a particular URL.

Zero matches

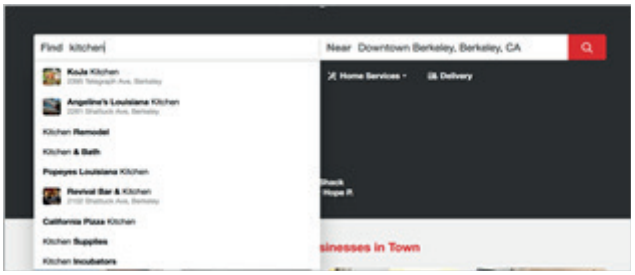
In some cases, there will simply be no matches for the letters typed. This can be a spelling problem, so as a fallback, use the search engine spellchecker, and if a very good match is found, display it as the only option. Users will generally select that correct spelling without confusion, as it is what they meant to type in the first place.

It may be a scope problem, for example, there are no relevant items on an iced dessert site for queries such as iphone or n95. This is a case where showing no suggestions gives customers negative but useful feedback.

High traffic zero matches terms should be marked as anomalies and domain experts can add them as internal searchable text, or block lists.

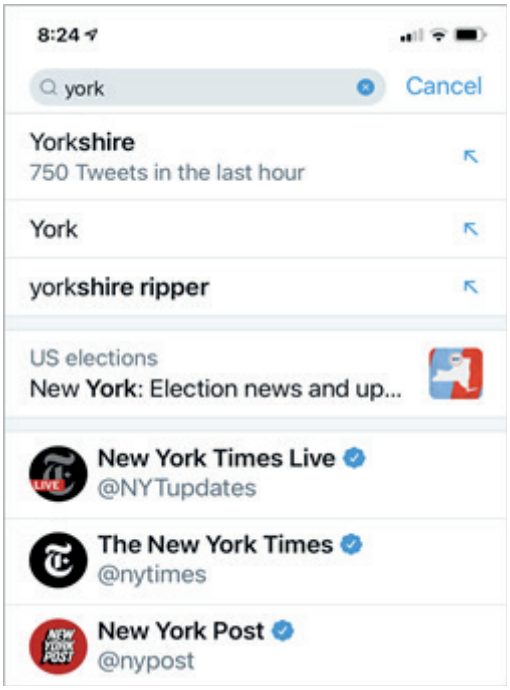
Suggestion menu user interface

In the display of suggestions, the most common user interface bolds the text that matches what users have already typed, and A/B testing tends to support that. However, some sites, such as Yelp, successfully present the suggestions with the term that matches in regular weight, and the remaining text bolded.



Auto-suggest at Yelp, with the match word "kitchen" in regular weight text

Images in the suggestions menu should be limited to controlled suggestions such as brand, type or very specific products. Even in ecommerce sites, it's difficult to create and curate small images that are sufficiently distinct to be recognisably different in this situation.



Auto-suggest at the New York Times showing the twitter profile images

Auto-suggest evaluation: an ongoing process

To measure and evaluate auto-suggest usage, the system should log a flag recording whether the menu was displayed, and if one of the entries was selected, and at what position. This can be correlated with ongoing results of quality testing. For A/B testing, the logging should also indicate which test setting was used. These results can also be compared to non-suggestion queries and (where available) SEO query metrics to identify whether the auto-suggestions are performing well.

The enterprise search user experience

Martin White

Enterprise search can trace its origins back to the early 1980s and for the last 40 years the research and development efforts have focused on improving relevancy, be it in the context of optimising precision or optimising recall. Much of the recent research using artificial intelligence and machine learning has the same objectives. Another recent development has been a significant increase in the number of connectors on offer to support federated search, with many vendors able to index 200 or more different applications. Most search vendor web sites are predominately about convincing customers that a solution to all indexing and query challenges is available off the shelf.

Enterprise search is different to all other enterprise database applications in that the user interacts with a computer to obtain a nominally best match to a query. This may take several sessions over an extended period of time as the user realises that the query needs to be recast, giving a new set of results to work through. The importance of the user interface in supporting this optimisation is critical to the success of a search session and yet very little attention has been made to the design and management of enterprise search user interfaces. Vendors are very keen to offer federated searching in multiple languages across both unstructured and structured content in almost any application or repository without explaining how the user is going to cope with a very complex user interface.

It does not help when the accepted wisdom (without any evidence) is that enterprise search should be intuitive and require no training.

We have reached the stage where all enterprise search software is based around a well-developed set of algorithms (for example BM25) and yet users are still dissatisfied with their organisation's search applications. In my experience the issue is not about what might be described as the technical performance of the search application but of the difficulty of using an interface which is almost always generic and not optimised for specific tasks. Research into professional search has shown that the ways in which lawyers, clinicians, patent agents and recruitment managers use the features of a search application are different. For example, clinicians make extensive use of synonyms (heart cf. cardiac) and acronyms, neither of which are of value to lawyers.

It does not matter how sophisticated the enterprise search technology is in terms of the features and functionality it offers. What matters is if the user can make an informed judgement about which piece of content presented in the list of results best serves their information requirement, reinforces their trust in the application and maintains the highest possible level of overall search satisfaction.

How the results are presented to the user is therefore critical in enabling the process of relevance judgement.

Result scanning

We tend to talk very glibly about scanning a list of results from a search without for one moment considering what this action involves. The speed at which the results can be scanned and appreciated in terms of their potential relevance varies from searcher to searcher. The concept of perceptual speed is usually totally ignored. This is a cognitive ability that determines the speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other tasks involving visual perception.

It is important not to confuse perceptual speed with readability. Perceptual speed relates to the ability of the searcher to make out words and other information elements. Readability is about the comprehension of those elements in the process of extracting information and knowledge.

Perceptual speed is not easy to measure but the impact on the search user can be quite dramatic. Users who are dyslexic face particular problems in scanning at speed. An outcome might be that the process of scanning is slow enough for a user to give up on the process after a few pages of results, and so not find all the relevant items.

Result reviewing

The next step is look at each result and decide whether it is relevant enough for us to take the time to click on it and open up the content item. Simple! Or is it? In common with many aspects of enterprise search there seems to be no research on how snippet length and design enable an informed decision on relevance to be made.

There is some (but arguably not enough) research on snippets for web search queries but in general these snippets are linking to a web page which can be scanned and assessed reasonably quickly. In enterprise search the content item could be several hundred pages long and it may be far from obvious where the relevant information (according to the ranking algorithm) is to be found.

There are three fundamental ways of generating a snippet:

- present the query term in a text sequence that should provide enough context for the relevance to be assessed
- create a computer-generated summary of the content item
- reproduce the first few lines of an abstract (see Google Scholar for examples).

Some search application vendors provide a thumbnail of a page that contains the query term, but the accessibility problems arising from having to view a small image displayed as the result of very precise mouse control are ignored.

The duality of search use

Enterprise search is almost certainly used by the majority of employees in an organisation. Most of the queries will be sub-critical, in that search is a convenient way of tracking down specific items of information but not the sole way of doing so. However there will be many employees who will be using the application to enable business-critical decisions to be made where a failure to locate the information needed could put the organisation at risk. It is also likely that in these situations the employee cannot be certain of which applications (and that is a deliberate plural) they need to search through in order first to find a range of relevant information and then have to integrate the results and synthesise an outcome to make as informed a decision as possible.

Over the last two years there has been a significant increase in the amount of academic research being undertaken into the search process. David Maxwell's thesis on information foraging and stopping distances makes a significant contribution to understanding the cognitive processes behind reviewing search results.

The screenshot below comes from research carried out by Hugo Huurdeman and his colleagues at the University of Amsterdam and the University of Nottingham. SearchAssist integrates both a results display and a range of search support features to support a multi-stage query and result process.

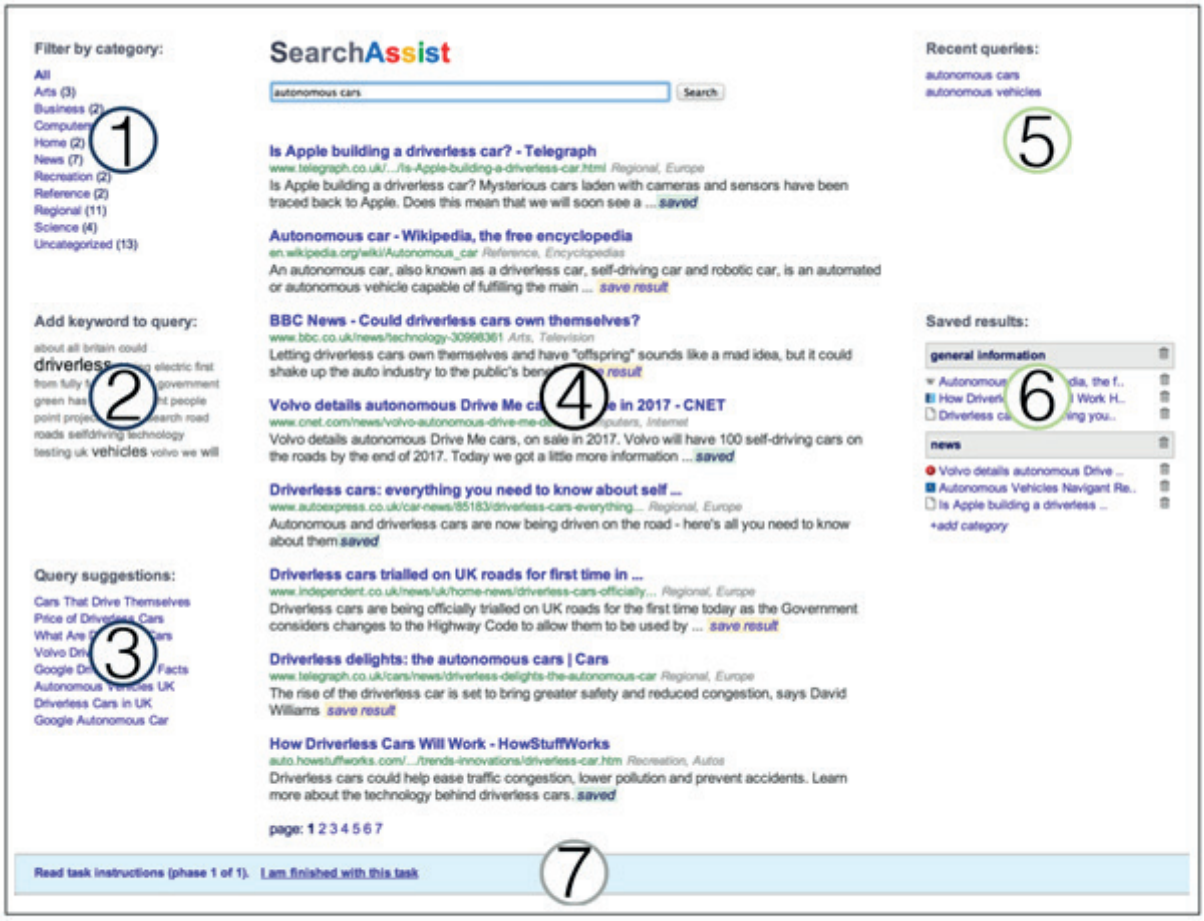


Figure 5.1: Screenshot SearchAssist. *Left column* (1, 2, 3): control features. *Middle* (4): input and informational features. *Right Column* (5, 6): personalizable features. (7): task bar

(From H.C. Huurdeman Supporting the complex dynamics of the information seeking process PhD thesis University of Amsterdam 2018 ISBN9789082169508 https://hdl.handle.net/11245.1/1e3bf31a-0833-4ead-a00c-4cb1399d0216.)

In addition there will be employees will be using the enterprise search application on a regular basis, perhaps several times a day, and will become conversant with even a very complex user interface. It is quite probable that they will wish to optimise the layout for different but related tasks. Research by Tony Russell-Rose, Jon Chamberlin and Leif Azzopardi into the search interface requirements of professional searchers indicates some important differences between the way in which patent agents, recruitment agents, lawyers and healthcare professionals use elements of the user interface.

Customising the user interface

The concept of customisable user interfaces has been under consideration for a number of years. It is well worth reading the outcomes of the Khresmoi project conducted in 2013/2014 with the objective of developing a multilingual multimodal search and access system for biomedical information and documents. This was achieved by:

- effective automated information extraction from biomedical documents, including improvements using manual annotation and active learning, and automated estimation of the level of trust and target user expertise
- automated analysis and indexing for medical images in 2D (X-Rays) and 3D (MRI, CT)
- linking information extracted from unstructured or semi-structured biomedical texts and images to structured information in knowledge bases
- support of cross-language search, including multilingual queries, and returning machine-translated pertinent excerpts
- adaptive user interfaces to assist in formulating queries and interacting with search results.

This research project led to the development of ContextFlow as a search application specifically designed for radiologists.

Sinequa, a French enterprise search software company, offers a range of customisable user interfaces, predominantly for the pharmaceutical and manufacturing sectors. The image below is a user interface for research scientists searching both internal and external sources of information. This interface can be customised by the search users.

© Sinequa. Used with permission.

A tipping point for enterprise search?

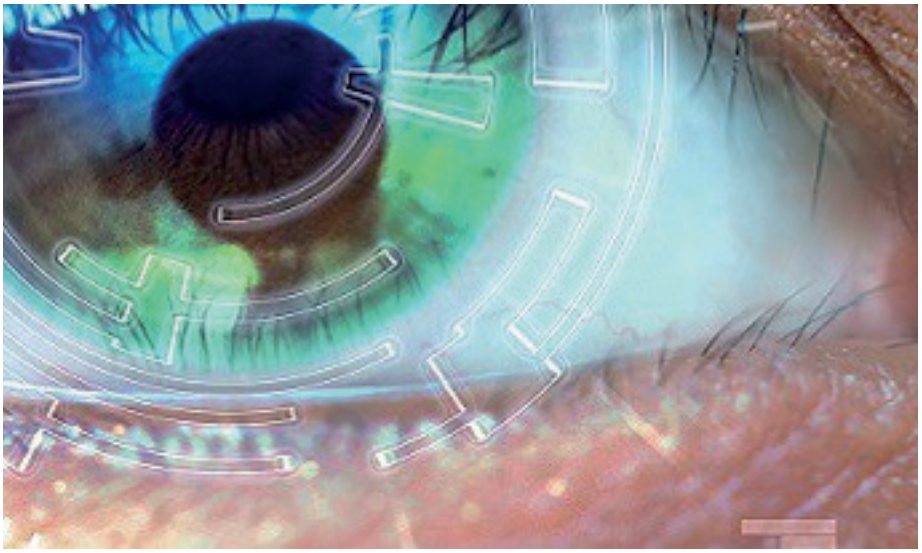
The technology behind enterprise search dates back to the early 1980s, with a major leap in functionality with the adoption of the BM25 ranking model in the mid-1990s. The BM25 model has gone through a number of variants and is now complemented with knowledge graphs and AI/ML routines. However the perceptual impact of these developments on users of enterprise search applications is arguably increasingly limited because of the inherent issues of the variability of content in enterprise collections and the range of intents of users. It is becoming very difficult for enterprise search vendors to differentiate their product offerings!

There now seems to be a substantial opportunity to offer search user interfaces that are optimised for specific tasks and/or capable of customisation by users. This approach is already being used by Sinequa in a number of areas, notably searching through clinical trials data and in product development applications. The impact of these developments on the search user is of course very visible, and the benefits in terms of productivity, innovation and speed of response to customers can be much more easily quantified than in text search. As a result it is easier for vendors to make a business case to prospective customers and to differentiate their offerings from competitors.

The result is likely to be that vendors and integrators will quickly appreciate the benefits of providing a much higher level of enterprise search user experience than they have over the last 40 years.

Searching fast and slow

Tony Russell-Rose, 2-D Search



Have you ever had that feeling of seeing something out of the corner of your eye, then turned to look but it's gone? We're left feeling cheated, as if some significant event has eluded our attention. But the reality is more prosaic: cells in the human retina are arranged so that movement and contrast are better perceived around the periphery, with the central region better suited to colour and detail. The result is that peripheral vision perceives things that the central region disregards. It's a simple explanation, but one that reminds us that in order to understand the bigger picture, we sometimes need to see in different ways.

In many ways, searching for information presents a similar challenge: in order to satisfy complex information needs, we must articulate those needs faithfully and then perceive their effect in the form of a response from the environment. We become partners in this exchange: a dialogue between user and search system that can be every bit as rich as human conversation. Crucially, the better we can articulate our own needs, the more trust we can place in the response.

Nowhere is this truer than for structured searching, where the goals of accuracy, transparency and reproducibility are at their most acute. In healthcare, for example, it is vitally important that all relevant sources of evidence be considered in developing policy, guidance and interventions. This is especially true during a global pandemic, and healthcare research needs to build on scientific evidence gathered in a systematic manner as part of its due diligence. Systematic literature reviews play a key role in this by synthesising the complex, incomplete and at times conflicting findings of biomedical research into a form that can readily inform healthcare decision making. And the cornerstone of systematic literature reviews is a systematic, structured search strategy.

To illustrate, let's take a familiar example: a complex search on the subject of *'Galactomannan detection for invasive aspergillosis in immunocompromised patients'*. In its traditional form, this would be articulated via a form-based query builder as a series of interconnected Boolean expressions:

```
1 "Aspergillus"[MeSH]
2 "Aspergillosis"[MeSH]
3 "Pulmonary Aspergillosis"[MeSH]
4 aspergill*[tiab]
5 fungal infection[tw]
6 (invasive[tiab] AND fungal[tiab])
7 1 OR 2 OR 3 OR 4 OR 5 OR 6
8 "Serology"[MeSH]
9 Serology"[MeSH]
10 (serology[tiab] OR serodiagnosis[tiab] OR serologic[tiab])
11 8 OR 9 OR 10
12 "Immunoassay"[MeSH]
13 (immunoassay[tiab] OR immunoassays[tiab])
14 (immuno assay[tiab] OR immuno assays[tiab])
15 (ELISA[tiab] OR ELISAs[tiab] OR EIA[tiab] OR EIAs[tiab])
16 immunosorbent[tiab]
17 12 OR 13 OR 14 OR 15 OR 16
18 Platelia[tw]
19 "Mannans"[MeSH]
20 galactomannan[tw]
21 18 OR 19 OR 20
22 11 OR 17 OR 21
23 7 AND 22
```

Each line consists of a series of keywords, operators and controlled vocabulary terms, which are connected via logical operators and Boolean expressions. The glue that binds all this together is the line numbering (a mechanism not entirely dissimilar to that used in early programming languages such as Unstructured BASIC).

Now, here is the test. If you were asked to describe how this search is structured, what would you say? How many conceptual elements does it contain? How are they related?

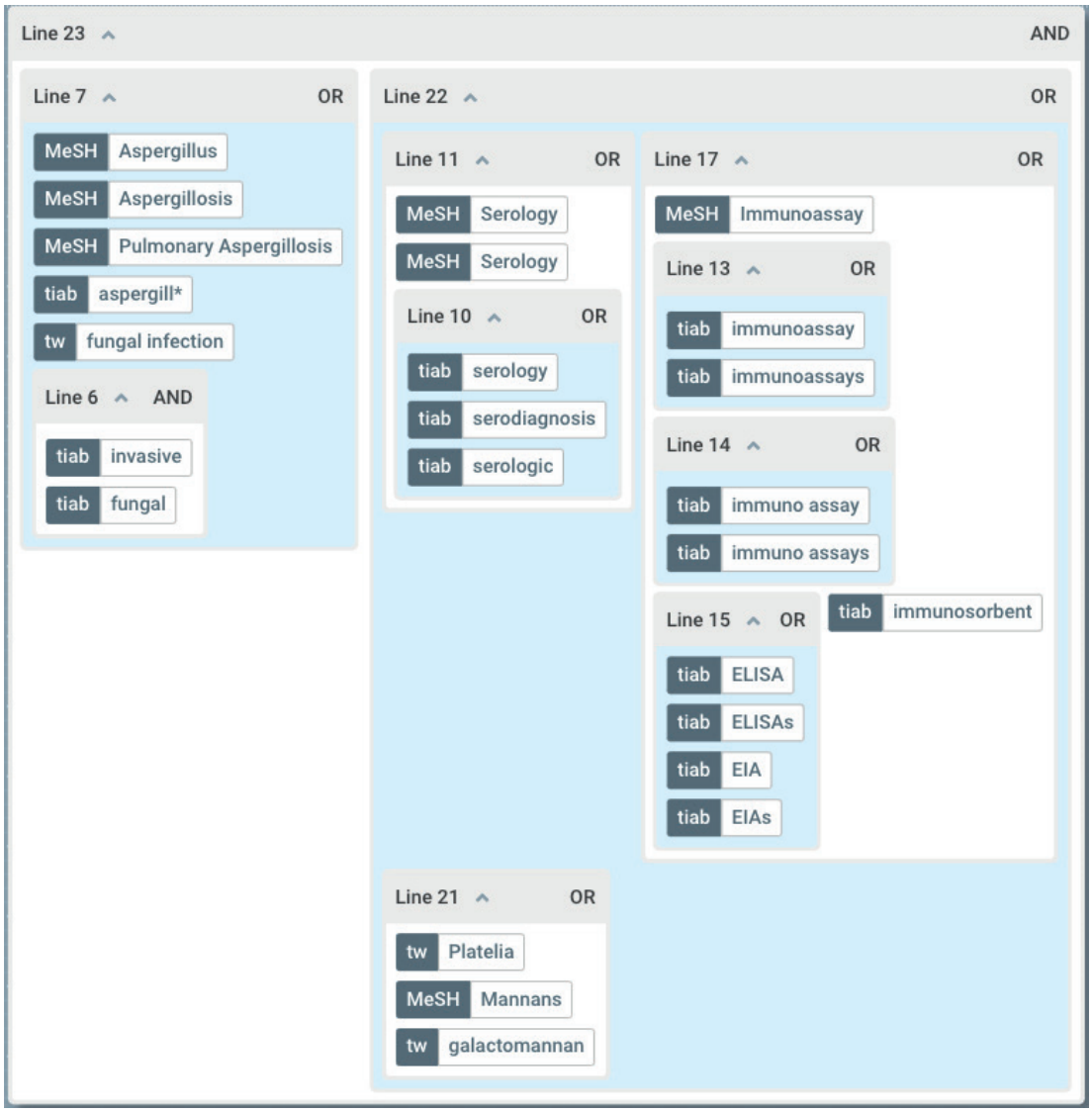
Clearly all these questions are answerable, albeit more so to the trained eye. But the point is that the answers are not directly visible. Instead, we must proceed through a sequence of steps: we must first retrieve from memory a method for interpreting line-by-line searches, and then implement it. In doing so we must hold data in our short-term memory, and keep track of where we are and where we are going, while holding on to any intermediate results. The process is mental work: deliberate, effortful, and laborious: a prototype of slow thinking.

And this is precisely where existing formalisms fall short. Just when we most need an effective way of seeing, we are left with words, lines and numbers. Instead of using perception to understand the structure of our information needs, we are forced to rely on cognition, with its associated human costs of effort and error. Instead of using approaches that allow us to think fast, we rely on formalisms that force us to think slow. Does it have to be this way? In what follows, we challenge this status quo.

Let's examine three alternative visions that are motivated by the principle of migrating thinking from slow to fast; from cognition to perception. We'll start with what we call the 'Nested view'. This view and those that follow can all be invoked by opening traditional, text-based search strategies using 2Dsearch.

Nested view

We've introduced this view before, so will review it only briefly here. In short, it provides a view which maps hierarchical structure onto a series of nested containers. The benefit is that the grouping and containment become immediately apparent:

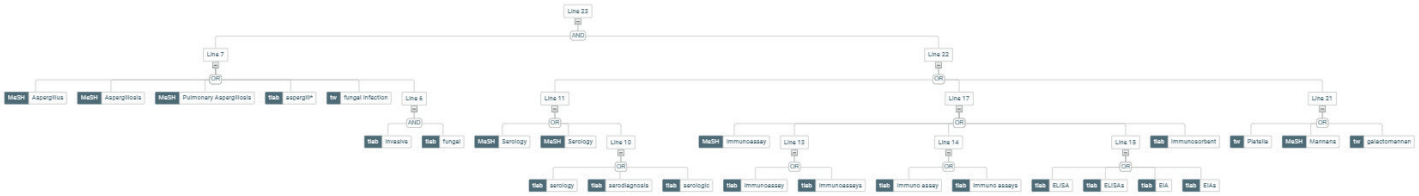


This visualisation reveals that our search strategy from earlier consists of a conjunction of two disjunctions (lines 7 and 22), the first of which articulates variations on the fungal infection concept, while the latter contains various nested disjunctions to capture the diagnostic test (serology) and associated procedures. By displaying them as nested groups with transparent structure, it offers support for abstraction, whereby lower-level details can be hidden on demand. In addition, it is now possible to give meaningful names to sub-elements, so that they can be re-used as modular components.

However, the Nested view has its drawbacks. Although it provides a fine degree of control over the arrangement and layout of the groups, the fact that they are rendered as blocks with operators in their headers isn't for everyone. Let's examine an alternative.

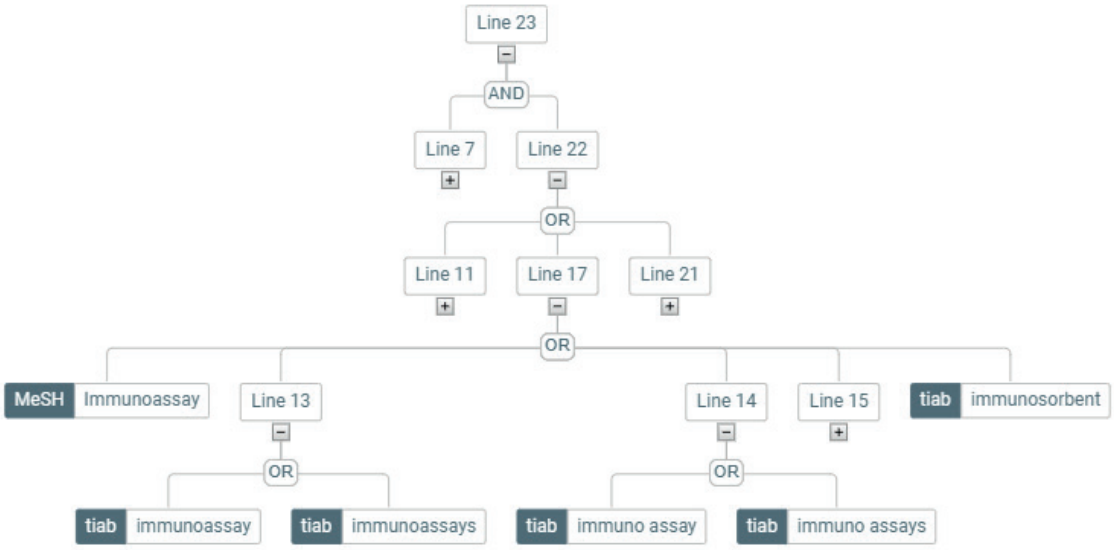
Tree view

Another way to understand the hierarchy embodied in complex searches is to apply a metaphor that is almost universally understood: the family tree. In this view, the search is represented as a visual hierarchy, with the root node (Line 23 in this example) at the top, and each level below represented as successive generations of children:



In this example, we have displayed the search in its entirety, and shrunk it to fit across the page. But it's easy enough to zoom in and out, and reveal just the higher levels:

Or to close branches on demand, and focus on one particular region of the tree:

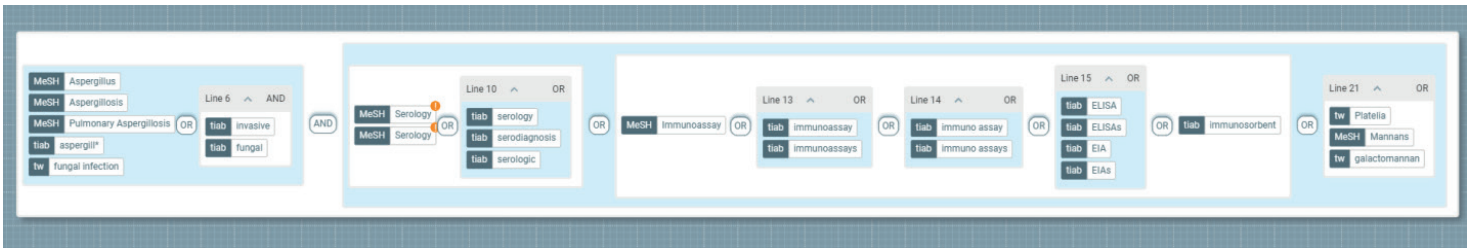


Like the nested view, the Tree view maps conceptual hierarchy onto physical hierarchy, but in a manner that emphasises branching over containment. But is hierarchy the most important aspect of a search? With that in mind let's examine a third view.

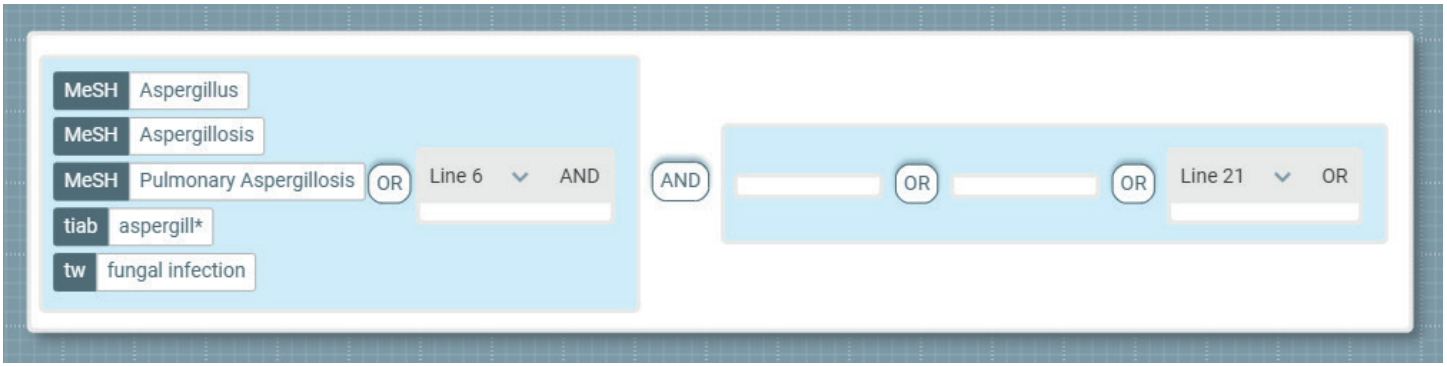
Inline view

The use of Boolean strings to represent complex searches may be inefficient and error-prone, but it does offer one key benefit: the ability to be read in a left to right manner. Of course, this attribute may reflect nothing more than the inertia of decades of convention, but there remains something useful about being able to read searches as a series of statements or commands. Is it possible to support this principle in a visualisation? This is where the Inline view comes in.

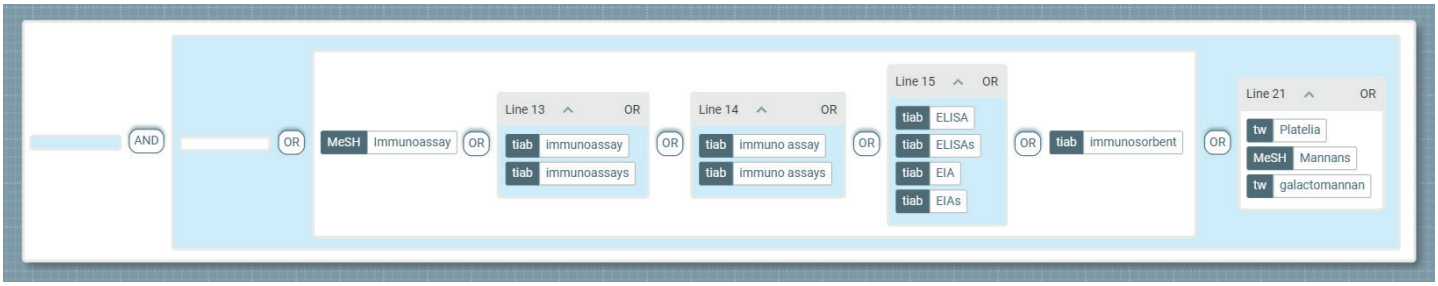
Like the Nested view, the Inline view maps conceptual hierarchy onto physical containment, but this time in a manner that aligns groups along a common midline, giving rise to a natural left-to-right reading:



Notice that in this view we elevate the operators to the same level as content items so that they appear in sequence within the left to right reading. This means that we can also reduce some of the chrome around groups, leading to a 'cleaner' layout. Again, we've shrunk the above image to fit across a single page. But as before it's trivially easy to zoom in and out, e.g. to reveal just the higher levels:



Or to close branches on demand, and focus on one particular region of the search:



Of course, the Inline view has its strengths and weaknesses too. However, it's important to recognise that while this article focuses on new ways of seeing, the real benefit is in the interaction: to modify a search, you can simply move terms from one block to another, or create new groups by combining terms. You can also cut, copy, delete, and lasso multiple objects. If you want to understand the effect of one block in isolation, you can execute it individually. Conversely, if you want to exclude one element, you can temporarily disable it.

In closing

In this chapter we've explored three different ways to visualise complex searches. In-each case, we've shown that it is possible to represent complex logic in a manner that supports both fast and slow thinking. Each view has its own strengths and weaknesses. Indeed, none of them is a silver bullet: the point is that they all reveal different aspects of a search strategy, and offer different insights and ways to understand them. It is through their collective diversity and flexibility that we discover new ways of seeing.
A picture is indeed worth a thousand words. To see for yourself, visit 2Dsearch, and let us know what you think.

Reinventing a neglected taxonomy

Helen Lippell

Introduction

In the 19th century British economist William Forster Lloyd coined the phrase "Tragedy of the commons" to illustrate the example of how unregulated grazing of livestock on common land by people acting only in their own immediate interest would result in the land becoming damaged for everyone. The concept of "tragedy of the commons" has since been applied to all sorts of areas such as overfishing, spam email, and toilet roll hoarding during a pandemic.

The following case study is a story of how taxonomies were central to an organisation's business model, yet had become unmanageable due to their not being anyone's responsibility. This situation damaged the quality of their products and was causing avoidable work and hassle for their staff.

Fortunately the organisation recognised that this had become unsustainable. As part of a wider programme to upgrade technical infrastructure that kicked off in late 2019, I undertook a review of everything related to taxonomies, including their structure, their management, and how they were being used to deliver information to paying customers. This article details how a data mess can be tackled both tactically and strategically, so that information-driven businesses can stop making life hard for themselves and their customers.

Background

This organisation has been around for decades. It built its reputation in the print era delivering high quality, unique information about the activities of government and the wider public sector to clients. Its digital products curate information and content from a wide range of sources including press releases, blogs, corporate websites, specialist publications, news media and proceedings from various parliaments in the UK and beyond.

Content is tagged with terms from a taxonomy (for UK information). Parliament-specific information is tagged with a separate set of controlled vocabularies and lists. Non-UK information is tagged with terms from a variety of small vocabularies. Another database stores custom queries, the rules by which the database identifies interesting information to send to clients. A typical custom query will contain a number of taxonomy terms relating to a single subject, and maybe also some free text keywords added by whoever created the query. When a new piece of content is tagged with a term that matches a term in a custom query, an email is sent to those clients who are interested in that subject. This system is designed to send the right information to the right people at the right time.

However, the taxonomy used for UK content had not been actively managed for a number of years. As a result, it had sprawled out of control. Not unlike the over-grazed common land, the taxonomy had become unusable for the people who were using it to tag content, as well as for the people who create and manage the custom queries. This was having a detrimental effect on the quality of the service to clients. Poor quality tagging was translating into inaccurate information being sent out.

There was a lack of documentation for staff to work from. No-one had been trained on the taxonomy, other than learning from other people who had been misusing the

taxonomy. There was no governance in place, and no style guide for adding new terms. The taxonomy was being added to in a haphazard manner. Taggers and the creators of the email alert queries were not working collaboratively.

The taxonomy

The taxonomy used by the UK part of the business was created some years ago by adopting the EuroVoc taxonomy. EuroVoc is managed by the EU Publications Office, and covers the activities of the EU. This was the first problem I identified - namely that the taxonomy was not built with UK-specific information in mind. It contained terms that are specific to EU bureaucracy, or expressed in 'Euro English' (the dialect used by non-native English speakers working inside and with EU institutions). The structure reflected the areas of interest to the EU, which overlap with, but are not the same as, how the UK public sector is organised. For example, health policy is not something which is traditionally co-ordinated at an EU-wide level, except for circumstances such as procuring vaccines. On the other hand, it is an area to which the UK government devotes a lot of budget and governance. For this reason EuroVoc is not an adequate vocabulary on its own for this important UK policy area.

The taxonomy had been added to in the years since it was first adopted, but the overall structure had never been reviewed. There was very little information about what had been added, when, by who and why. It was impossible to see which terms were being used by the taggers (other than by checking individual pieces of content). It was also impossible to see which terms were being used in the email rules. There were tens of thousands of rules in the database; far more than the team could ever keep track of.

Assessing needs and taking action

Technology

The organisation had recognised that this situation was unsustainable, especially as it had started to affect the quality of service to customers. A wide-ranging review project had come up with a new technical architecture featuring replacements for numerous legacy systems.

Unfortunately the taxonomy was not initially included in the plan. There was an assumption that the taxonomy would just be managed in the back end of a content management system, rather than being stored in a fully-featured taxonomy tool.

It was important to work alongside the offshore team that was scoping and building the new infrastructure. The team was aware of the problems with the taxonomy and at one point had even proposed eliminating it altogether. Machine learning would provide all classification and hence all information delivery to clients would depend on rules curated from the automated tagging. The reality is, however, that manual tagging would still be needed, in order to deal with specialist areas of content and to ensure high quality for customers.

People

I talked to the teams who tag each piece of content as it comes through the system. They understood the need for tagging; after all, if content was only retrievable through free-text searching, it would be even harder to attain the level of quality that customers want. However, few of the team had ever had meaningful training on tagging accurately, or even on understanding what was in the taxonomy. They had developed a number of workarounds and unwritten 'rules of thumb' such as "if you see content about x topic, use y tag", even if that tag was inaccurate. (Customers would never see which tags

had been applied to the content they had been sent. But consistently using the wrong tag perpetuates a loss of understanding of what tagging is for, and what correct tagging should look like.)

Process

It was imperative to recommend that documentation and processes relating to tagging were created. There were three actions. Firstly, I set up shared documents containing definitions of commonly used and misused tags. These had the advantage of hopefully contributing to a common understanding across the team, as well as flushing out tags which people were struggling to use correctly. The shared resources became doubly important as a new offshore team started working on tagging.

Secondly, I proposed a basic process around governance for adding new tags. This is tricky to maintain in the absence of a permanent taxonomy manager. However, the tagging teams can share ownership of the taxonomy's future quality. Finally, I recommended new guidance on tagging correctly, and on working more closely with the email rules team.

The taxonomy

I gathered examples of inaccurate tagging and misleading tag definitions to highlight the problem for senior management. I encouraged managers to view the taxonomy as a core piece of what made the business distinctive compared to other websites and services that deal with current affairs and government policy.

This helped shift their expectations about the focus of the later taxonomy review. Instead of merely cleaning up a mess, the review would look at the taxonomy in its entirety, from the high-level structure to the individual terms. It would be reshaped around the organisation's expert and specialised knowledge of the UK domain. It would contain concepts and language that was used by customers and in content.

Outcome and learnings

The most important outcome of the work was not technical but human - the acknowledgement by senior management that the taxonomy was a valuable business asset. Their customers don't buy information services from them because of the technical infrastructure (as good as that will be once the project is complete). Rather, they buy information because it is accurate, useful and delivered quickly. The taxonomy is a key part of their intellectual property (IP), and once the remediation work is done, it will reflect much more closely their knowledge of both the domain and their customers' range of interests.

They have now hired an experienced taxonomist to do the taxonomy review. They will ensure that staff, including technology and editorial, are trained to understand and use the taxonomy properly. For the taxonomy to be sustainable, it is critical that the organisation embeds taxonomy and search skills for the long-term. This will minimise the risk of getting into a similar situation in the future. Much like William Forster Lloyd's parable of overgrazed land, a taxonomy needs to be managed with regard to the bigger picture of sustainability, as well as supporting the needs of its individual users.

Ecommerce site search is broken: how to fix it with open source software

Charlie Hull & Eric Pugh, OpenSource Connections

Imagine the scene: you've walked into a store intending to buy a coffee maker. It's the end of the week, you're tired, you want to finish your shopping and get home. After a few minutes wandering the aisles you approach a sales assistant and ask them for help. Here's how the conversation goes:

Customer: "I need a coffee maker"
Assistant: "Sure, here's 305 things you could buy all related to coffee."
Customer: "No, I said coffee maker, not ground coffee, coffee beans or coffee-flavoured chocolate. Try again, I want a coffee maker from the kitchen accessories department"
Assistant: "Sure, here's 55 things from that department related to coffee makers."
Customer: "You're misunderstanding me! I want one of those glass things with the bit you press down - cafe tear I think they're called. I can't see any of those in your list"
Assistant: "I'm sorry we don't have any cafe tears."
Customer: "Sorry, perhaps it's my accent, do you have anything that sounds like 'cafe tear'?"
Assistant: "We have one thing, it's called a 'cafetiere', here it is."
Customer: "That's right! But the one you're showing me is for 12 cups, I want a smaller one with a metal handle. You're a big store, surely you sell more than one type? Look again for cafetieres?"
Assistant: "We actually have 12 different cafetieres - here they all are."
Customer: "Finally! Why on earth didn't you show me these just now, or when I first asked for a coffee maker? They're only used for making coffee! I nearly walked out of this store, you're lucky I'm still here..."

This conversation is based on some actual interactions with a real ecommerce site search engine on a major UK supermarket website. It illustrates some common problems: the customer using different terms to the merchant, over-long or over-short result lists, no automatic phrase boosting or spelling suggestions and eventually a frustrated customer who very nearly goes to a competitor. If a real-world sales assistant behaved like this they probably wouldn't last long in the job!

Ecommerce is now a vital lifeline for many people and a major business driver - a study in 2020 from Emarketer showed that "UK retail ecommerce sales will account for 27.5% of total retail sales this year, and that proportion will approach one-third by 2024". The COVID-19 pandemic has hugely accelerated a shift that was already underway and businesses who don't provide effective tools for ecommerce - including site search, the equivalent of a sales assistant - are at huge risk of loss of sales and brand reputation. Online consumers are fickle and it's far easier to switch to a different website than it is to walk across town to a different store - and often that website is Amazon or another giant competitor.

In short, ecommerce site search is broken. The reasons for this are manifold: the search engines provided by commercial ecommerce software are often badly integrated, out of date, hard to control and provide no way to measure search quality. Marketers, who best understand how to match customer needs to inventory - they know a cafetiere is a coffee maker and how many cafetiere types you sell - are seldom provided with the tools to influence or tune search results, or even told much about how the search

engine works. IT, tasked with keeping the lights on, may not be aware of business objectives or targets and thus find it hard to prioritise search-related issues. Lastly, divining the actual intent of a customer from a two-word phrase is difficult if not impossible.

There are obvious benefits of improving site search - more successful searches lead directly to more conversions and thus revenue - but there are other benefits. An ability to examine site search logs and other pointers to user behaviour may reveal those items customers are searching for that a merchant doesn't provide, a pointer to expanding inventory or to new trends and needs. Who would have predicted in 2019 that facemasks and hand sanitiser would need to be so widely available in 2020?

Our approach at OpenSource Connections (OSC) to improving search can be summed up as 'measure, experiment, repeat'. The first step is developing effective measurements of search quality - you need to know how bad (or good) search results are, and you must be able to measure this on a repeatable and frequent basis. The second step is to be able to easily make changes to search engine configuration and to assess the impact of these changes - the ability to experiment, rapidly and safely, offline. Once an offline experiment shows measurable improvements it can be promoted to online where A/B testing and click logs can be used to further measure impact.

This culture of rapid experimentation must be developed across the whole search team - not just within IT. We need to provide tools that marketers can use to react to rapidly changing situations, but we also need to base our testing on solid data. Our tools also need to be widely available, not tied to a particular platform or technology stack, well documented and battle tested. We need to give full control of search back to the people who need it.

OSC has been working with a number of others across the industry to bring together a suite of freely available, open source tools that can be used to build measurable and tunable ecommerce site search. The group has christened this initiative Chorus and based the development on one of the two leading open source search engines, Apache Solr, which is widely used in ecommerce, sometimes as part of commercial packages. A variant for Elasticsearch, the other popular engine, is in active development.

OSC's Quepid tool is one part of the ensemble, allowing one to create test cases, add queries to those cases and collaborate with subject matter experts to give judgements of search quality. Quepid lets users (who need no deep search expertise) 'rate' search results on a scale using a simple web interface and gives an overall quality score. Importantly once a change has been made to the search engine configuration, Quepid can easily re-run a test and the change to the overall quality is shown, allowing promising experiments to be identified.

Another part of the suite allows business rules to be added directly to the search engine, for example synonyms, to help with the problem that your customers may not use the same language as you do when describing products. Boosting is another technique available to move certain results higher up the list. It is also possible to turn dimensions into ranges - for example, to match up a customer looking for a 33-inch TV screen and a merchant who sells 32-inch and 36-inch screens, both of which may be acceptable results as they are close in size. Querqy is a query preprocessor that helps turn this customer language into an effective search query, and SMUI is a web interface that helps manage these business rules. These two tools give search teams improved capabilities in active search management, also known as 'searchandising'.

Let's return to our example above: how might you fix it with Chorus? First, we would use our search logs to make sure we were testing common search queries: if 'coffee maker' was a common query then a test should be run for it (if not, perhaps there are more important things for our team to consider given limited resources and time). We would then use Quepid to create a test case including the query 'coffee maker' and ask our subject matter experts - our marketers - to rate the results. Using this ground truth data we would try some experiments to see if we could improve things: perhaps 'cafe-tiere' (or 'french press') should be created as a synonym for 'coffee maker', or a boost applied if the result was in the 'kitchen accessories' category. We can try both these techniques using SMUI and Querqy and rapidly see how results are affected. With Solr or Elasticsearch we can also try different spelling suggester configurations which might help with 'cafe tear'. Eventually once our offline testing had identified some candidate improvements we would consider this for online testing.

There are several other components which can assist with large scale batch testing, finding optimum configuration parameters and automated deployment of the platform. Our group is already working with a number of leading ecommerce websites to deploy Chorus and give control of site search back to search teams. We also welcome any contributions to the project.

Come and join the Chorus!

Further reading

1. UK Ecommerce 2020 - Digital Buying Takes Hold as Pandemic Decimates the High Street https://www.emarketer.com/content/uk-ecommerce-2020
2. Test your site search with a free downloadable assessment guide www.opensource-connections.com/guide/ecommerce
3. Meet Pete the Product Owner - a series of blogs and videos demonstrating Chorus https://opensourceconnections.com/blog/2020/07/07/meet-pete-the-e-commerce-search-product-manager/
4. https://github.com/querqy/chorus to download Chorus
5. www.querqy.org for Chorus documentation
6. Join the free search community Relevance Slack at www.opensourceconnections.com/slack

The data-driven search team

Max Irwin, OpenSource Connections

High search quality is a necessary aspect of any successful business. What can you learn from the teams that have mastered this capability that you can replicate and use successfully in your business?

The well-functioning search team does two things: its members (1) rely on understanding customer needs based on how they express themselves through the product experience, and (2) leverage this data for improvement and decisions throughout the organisation.

In this article I will describe the need for data across the organisation, and how the various roles in a search team enable and consume that data. We will talk about roles and responsibilities across the organisation, grouped by competencies of the search team.

Decisions from data

"We need to be a data-driven organisation" may sound like a cliché but that doesn't make it untrue. The most successful companies rely on gaining insights from expressive customer data and improving the product using those insights.

Search provides the purest form of customer expression for most products: freeform text in the form of queries, paired with the context of the person executing these queries. Most other forms of behavioural data that don't include text, such as UI A/B tests, navigation and workflow paths, and other types of engagement metrics, can provide detailed insight into how the customer uses the product as it exists. Having a text box that anyone can type anything into and expect results, is the beautiful chaos of what your customers really want.

The only other expression tool that approaches search is customer feedback. Direct (feedback forms, surveys, and emails) or indirect (social media mentions) are included, but these types of feedback are only done by a small subset of the customer base and are usually lopsided towards negative feedback. The vocal minority should not be the only set of customers to drive changes to the product and/or its features.

At OSC we use search as the key to understanding customers, and react according to the data the search product yields.

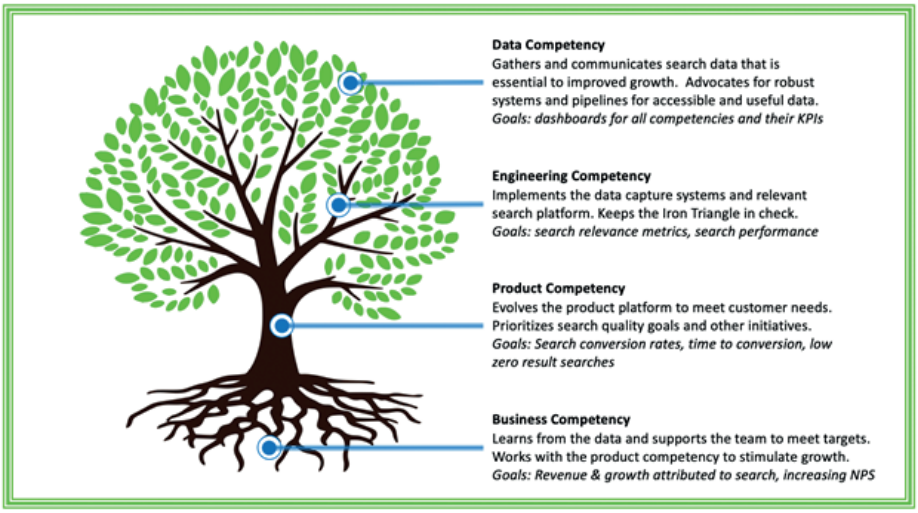


Figure: The data competency tree

Business competency

How can the executive, sales, and marketing group guide search quality, without understanding how or why it works? The good news is that not much changes here aside from awareness and alignment. KPIs related to revenue, profitability and customer growth are still the north star for the business, and these lagging KPIs are influenced by search engagement. The trick is attributing these lagging metrics to search.

Depending on the domain, it may be difficult to arrive at these numbers. For eCommerce and asset search, it is straightforward. But how does one understand whether search is not only successful, but successful to the point of raising the business?

For eCommerce, I wrote extensively on site-search KPIs in a series of blog posts: (https://opensourceconnections.com/blog/2020/08/28/e-commerce-site-search-kpis/).

For other types of search, I recommend focusing on Net Promoter Score (NPS) while directly asking about search during the survey. Correlate this with data based on engagement.

Roles and their impact on search

Follow these team responsibility guidelines to enable a well-performing organisation that leverages search data.

- **Executive Stakeholder** – this is usually an SVP or VP of technology for medium/large businesses, and the CTO for small business. Sets revenue/growth targets and coordinates the executive team to rally around improvement. Drives the main strategic decision for search quality with the product owner. The executive stakeholder must also understand the relationship between leading product metrics of search quality and their impact on the lagging metrics of revenue and growth.
- **Marketing Officer** – ties site-search and content improvement with organic search (SEO) and campaigns. Works with the Engineering team to evangelise search technology internally and with the wider community to grow the technical brand of the search team.
- **Account Manager** - works with client/customer representatives and agencies to ensure high priority search quality feedback gets to the Product Owner.
- **Product Owner** - straddles the Business and Product teams to ensure unified improvement for the search product. Expresses business goals as OKRs to the product team. Ultimately accountable for all search quality initiatives.

Product and customer experience competency

Moving to finer detail into what makes up the high-level business goals of revenue, we look to identify content/item and feature level impact on the customer, and how it shapes the product. The goal for the product team is to increase customer satisfaction. Higher satisfaction will correlate with a higher revenue/growth metric for the business. Since revenue and growth are lagging metrics, we focus on leading metrics for the product as an immediate marker for predicting this business improvement. For this, we use the KPIs of search conversion rate, time to conversion, and search abandonment/exit rate. These metrics can be set as KPIs in the team (with reasonable targets for improvement over time). Zero/low result searches are also an important metric to use as they identify gaps in the vocabulary and/or content.

Roles and their impact on search

Follow these team responsibility guidelines to enable a well-performing product that continually evolves to meet customers' ever-changing needs.

- **Product/Project Manager** - decides goals and roadmap for the product with the Product Owner. Keeps the product and engineering teams on schedule and unified towards key search measurement, experimentation, testing, and deployment. Reacts to trends and changes that customers expect, facilitated through hypothesis-driven-development as an Agile methodology.
- **Business Analyst** - expresses goals as search requirements and stories. Uses dashboards and reports to inform and influence the product and content direction. Responsible for capturing explicit feedback (judgement lists) for search quality measurement and understands impact of changes made by the engineering team.
- **Content Owner** - responsible for creation/curation and quality of content or product metadata for search. Includes merchandising, landing pages, content style, and vocabularies.
- **Customer Experience** - defines personas and customer information needs. Accountable for wireframes and behavioural flows for search, results, facets, autocomplete, spelling, highlighting, and all other areas of the search experience.
- **Support** - aids customers having trouble with search and informs the Product and Engineering competencies of areas of improvement to increase satisfaction and reduce support/helpdesk overhead.
- **UI/Design** - expresses the search experience wireframes/workflows as implementable designs. Responsible for brand and style conformity, accessibility, and attractiveness of the product. Some teams may include front-end development in this role.

Engineering competency

Moving into the technical aspect of search, the Engineering competency supports the product with efficiency and relevance, expressed by formulaic measurements. Efficiency is set by performance KPIs and the SLA, Relevance metrics may include Normalised Discounted Cumulative Gain (nDCG), Mean Reciprocal Rank (MRR), Expected Reciprocal Rank (ERR), and others.

Engineering must also support development of the search product as decided with the product competency, based on strategic direction from the business team. Metrics come from implicit or explicit judgements, with the latter usually provided by the product team in a tool such as Quepid (https://quepid.com). Implicit judgements are gathered from usage logs and verified by the data team.

Roles and their impact on search

Follow these team responsibility guidelines to build a snappy and relevant search system by leveraging the data feedback loop.

- Architect/Technical Leads - responsible for overall services, content and data flows, and system design in a multi-cloud environment. Ultimately accountable for design, improvements, and quality of the search platform. Designs search service, API contracts, and analytics capture touchpoints. Coordinates search efforts across the entire group.
- Developers - implement the services as agreed with the architect and DevOps roles. Ultimately accountable for shipping working code of a high standard.
- DevOps - responsible for CI/CD of the search platform, and sizing of the search engine and its infrastructure. Ultimately accountable for uptime, deployments, scalability, redundancy, backups, and performance.

- Quality Assurance Engineers - responsible for testing all technical aspects of the search platform and whether it meets product team requirements. Implements test tooling, and automation for relevance and code quality. Works closely with the rest of the competency and design to implement A/B or multi-armed-bandit testing frameworks. Identifies gates and safeguards for the CI/CD of the platform.
- Relevance Engineers - responsible for the relevance of search results given the customer context and query. Develops and tunes all search features such as query parsing, result scoring, autocomplete, spellcheck, highlighting, and others. Ultimately accountable for increasing nDCG or other search metrics. Informs Architect and DevOps to ensure performance of the search engine remains high.

Data competency

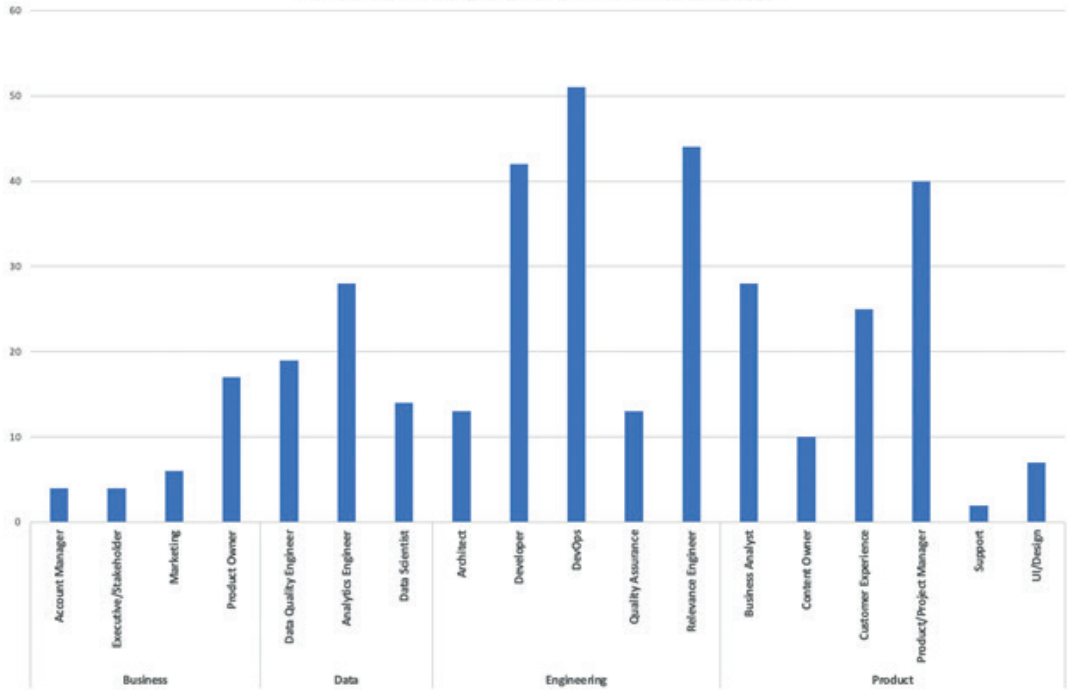
Can you correlate the lagging revenue and growth metrics with nDCG or search conversions? All of the above must be supported by data, and here is where the foundation for understanding is created and grown.

- Data Quality Engineers - responsible for automation and tooling of content and metadata quality. Works closely with the content owner.
- Analytics Engineers - responsible for implementation of data capture and reporting services. Works closely with the Architect, Developer, and DevOps roles to integrate a stack-wide data platform that captures customer behavior and search platform performance. Identifies gaps in metrics and logs and works to fill them over time based on priority.
- Data Scientists - responsible for converting raw analytical data into actionable reports and insights. Develops and tunes models for Learning-to-Rank and Natural Language Search.

Roles and responsibilities of a search team

We surveyed several companies to understand what roles make up their successful teams and what their responsibilities are. This chart gives an overview of the involvement of each competency.

Search Team Responsibilities by Group and Role



In conclusion

Let's put it all together! Here is the cycle of understanding:

- The **product competency** expresses known customer information needs and personas that are a priority for the business. The product competency also curates explicit feedback to be used by engineering in baseline setup.
- The **engineering competency** implements baseline relevance and systems, and integrates analytics and data capture points based on direction from the data team.
- Once live, search data is captured and is used by the data competency for analysis, a and to create reports and dashboards showing the important metrics that all competencies require.
- The **product, engineering**, and **data competencies** iterate to optimise their KPIs over time.
- As search conversion rates increase (and abandonment decreases), customer satisfaction increases. Customer satisfaction is then seen as increasing the lagging KPIs of revenue, customer growth, and NPS.

Having these activities in place is a necessary condition for the of improvement of search quality. A highly capable data driven team will accelerate the business to success.

Search resources

The books listed below represent a core library which should be on the bookshelf of any manager with enterprise search responsibilities. They are listed in reverse chronological order.

Trustworthy Online Controlled Experiments
Ron Kohavi, Diane Tang and Ya Xu (2020), Cambridge University Press (Review)
An excellent introduction to A/B testing, which is a corner-stone of information retrieval evaluation.

Systematic Searching – Practical Ideas for Improving Results
Paul Levy and Jenny Craven (editors) (2019), Facet Publishing (Review)
A core technique in undertaking systematic reviews, with wider implications for high-recall search.

Understanding and Improving Information Search – a Cognitive Approach
Wai Tat Fu and Herre van Oostendorp (co-editors) (2019). Springer (Review)
A collection of papers looking at information retrieval performance from a cognitive perspective.

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.

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.

Looking for Information
Donald O. Case and Lisa M. Given (2016, 4th Edition), 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.

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.

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.

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.

Enterprise Search
Martin White (2015, 2nd Edition), 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.

Designing the Search Experience
Tony Russell-Rose and Tyler Tate (2012), Book website) (Review)
This book takes a deeper look into information seeking models, using them to consider how best to design user interfaces.

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.

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.

Morgan Claypool, Now Publishers and Manning Publications offer a wide range of books on specialist aspects of information retrieval and search. The books from Manning Publications are written specifically for search developers and search managers. The books from Morgan Claypool and Now Publishers have more of an information retrieval focus.

This is a list of experts who write about and comment on aspects of search technology and implementation on a reasonably frequent basis.

Beyond Search Stephen Arnold
Complex Discovery Rob Robinson
Coveo Insights Corporate Blog
Daniel Tunkelang
Do More With Search BA Insight corporate blog
Elastic Corporate blog
Enterprise Search Professionals (LinkedIn)
Geodyssey Paul H Cleverly
Information Interaction Tony Russell-Rose
Intranet Focus Martin White
LucidWorks Corporate blog
Opensource Connections Corporate blog

Searchblox Corporate blog
Search and Content Analytics Blog Paul Nelson, Search Technologies
Search Explained Agnes Molnar
Sease Corporate blog
Sinequa Corporate blog
Synaptica Corporate blog
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.

There are two regular columns on search published by CMSWire:

https://www.cmswire.com/author/miles-kehoe/
https://www.cmswire.com/author/martin-white/

List of enterprise search vendors

Company	HQ	URL
Algolia	USA	https://www.algolia.com
Amazon	Denmark	https://aws.amazon.com/kendra/
Ankiro	USA	https://ankiro.dk/ankiro-enterprise-suite/
Aras	USA	https://www.aras.com/en/capabilities/enterprise-search
BAInsight	USA	http://www.bainsight.com
Bloomreach	USA	https://www.bloomreach.com/en
Bonsai	USA	https://bonsai.io/
Cludo	Denmark	www.cludo.com
Cognite	Norway	https://www.cognite.com/en/company
Copernic	Canada	http://www.copernic.com/en/products/enterprise-search-engine/
Coveo	USA	http://www.coveo.com
Curiosity		https://curiosity.ai/
Datafari	France	https://www.datafari.com/en/
DieselPoint	USA	http://dieselpoint.com
Dokoni Find	USA	https://www.konicaminolta.co.uk/en-gb/software/dokoni-find
dTSearch	USA	http://www.dtsearch.com/
Elastic	Netherlands	https://www.elastic.co/elasticsearch
Exalead	France	https://www.3ds.com/products-services/exalead/products/
Expert.ai	Italy	https://www.expert.ai/
Fess		https://fess.codelibs.org/
Findwise	Sweden	http://www.findwise.com/en
Funnelback	Australia	https://www.funnelback.com/
Gimmal	USA	https://www.gimmal.com/records-management/enterprise-search
Google	USA	https://cloud.google.com/products/search/
Grazitti		See SearchUnify
Hulbee		https://hesbox.com/en/overview/at-glance
Hitachi	USA	https://www.hitachivantara.com/en-us/products/storage/object-storage/content-intelligence.html
Hyland	USA	http://www.hyland.com/en/products/enterprise-search
IBM Watson	USA	https://www.ibm.com/watson/products-services
Ilves	Finland	https://ilveshaku.fi/en/
iManage	UK	https://imanage.com/
Inbenta	Spain	https://www.inbenta.com/
Indica	Netherlands	https://indicaplatform.com
Infongen	USA	https://www.infongen.com/solutions/enterprise-search
Intergator	Germany	https://www.intergator.de/en/solutions-applications/enterprise-search/
IntraFind	Germany	https://www.intrafind.de/index_en
Klera	USA	https://klera.io/
Knowliah	Belgium	https://www.knowliah.com/
Lucene	Community	https://lucene.apache.org/
Lucidworks	USA	http://www.lucidworks.com

Company	HQ	URL
Microfocus	UK	https://www.microfocus.com/en-us/products/
Microsoft SharePoint	USA	https://docs.microsoft.com/en-us/sharepoint/dev/general-development/search-in-sharepoint
Mindbreeze	Austria	http://www.mindbreeze.com
Nalytics	UK	https://www.nalytics.com/
Keeeb	USA	https://www.keeeb.com/
Netowl	USA	https://www.netowl.com/enterprise-search
Onna	USA/Spain	https://onna.com/enterprise-search/
Open Source Connections	USA	http://opensourceconnections.com/
OpenText	Canada	https://www.opentext.com/what-we-do/products/discovery
Precognox	Hungary	https://www.precognox.com/intelligent-search/
ResoluteAI	USA	https://www.resolute.ai/about
SAP	USA	https://blogs.sap.com/2019/08/16/enterprise-search-the-new-user-experience-for-enterprise-information-processing/
Searchblox	USA	https://www.searchblox.com/
Searchunify	USA	https://www.searchunify.com/
Sinequa	France	http://www.sinequa.com
Solr	Community	http://lucene.apache.org/solr/
Squirro	Switzerland	https://squirro.com/
Swiftype	USA	https://swiftype.com/
Tantivy	Community	https://github.com/tantivy-search/tantivy
Terrier	UK	http://terrier.org/
Theum	Germany	https://www.theum.com/cognitive-search-and-knowledge-discovery/
Thunderstone	USA	https://www.thunderstone.com/lp/enterprise-search/
Vespa	Community	http://vespa.ai/
Voyager	USA	http://www.voyagersearch.com

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 access permissions at a user or group level (often based on Active Directory) to 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 retrieving results, often using Boolean operators.

Aggregated search
The presentation of related content items (often referred to as verticals) from a single index in a specific area of a page of search results.

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.

Artificial intelligence
A set of technologies that enable machines to sense, comprehend, act and learn in a manner that seeks to emulate a human response to a situation.

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.

BERT
Bidirectional Encoder Representations from Transformers (BERT) is a machine learning technique which enhances the performance of training based on natural language processing.

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 (best match 25)
A ranking algorithm developed in the 1990s of which there are now multiple variants. It has its origins in the tf.idf ranking function and is widely used as the basis for enterprise search applications.

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 higher in the results than the raw algorithm would suggest.

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

Chatbot
An application that can conduct a voice query against a search index in lieu of providing direct contact with (for example) a call-centre operator.

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.

Conversational search
Search applications that respond to a spoken request or query with a spoken response. (See also Chatbot)

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.

Deep learning
Deep learning builds on machine learning principles but makes use of artificial neural networks to be able to manage very large collections of data with real-time responses.

Description
A brief summary, often generated automatically, that provides 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 an information-based application such as a content management system or enterprise search.

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.

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”.

Exploratory search
Search where the goal is imprecise and open-ended and there is no unique single answer that meets the user’s information needs and so it is difficult to determine when to terminate the search.

Facet
Presentation of topic categories and content metadata on the search user interface to support the refinement of a search query generated by the search index as the process of query exploration proceeds.

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

Federated search
A search carried out across multiple repositories, indexes 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 offers specific criteria for search result selection that is independent of the query e.g., file format or publication date.

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 queries and documents already marked as relevant by topic experts, 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 created as an outcome of a crawl of every word, entity and associated meta-data in a way that facilitates the very fast retrieval of documents.

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.

Knowledge graph
A representation of entities and related attributes.

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)

Learning to rank (LTR)
A class of techniques that apply supervised machine learning to solve ranking problems by presenting a relative reordering of relevant items.

Lemmatisation
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.

Long tail
A feature of text-based search in which there are a significant number of low-use queries forming a long tail which is difficult to optimise for an individual query. An example of a Zipf curve.

Machine learning
A method of data analysis that automates analytical model building.

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 supplements and/or clarifies index terms generated by text in the document, for example the date of publication or the author or specific controlled terms.

Morphologic analysis
The analysis of the structure of language.

Natural language processing
A process that identifies content through using grammatical and semantic rules to understand the intent of a sequence of words in a specified language.

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

Neural IR
Neural ranking models for information retrieval (IR) use shallow or deep neural networks to rank search results in response to a query.

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.

Professional search
A term applied to the way in which groups of professionals (for example lawyers and patent agents) develop complex queries in order to achieve very high levels of recall.

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
Search applications calculate a relevance score for each content item and return results in decreasing order of relevance.

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 query. Sometimes incorrectly referred to as 'keywords'.

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.

Session
The duration of the time spent by a user between entering a query term, reviewing results and then closing down the application.

Snippet
The text that is presented to give a concise representation of the content of a search result sufficient for a user to assess its relevance to their query. It may be generated by the author of the document, extracted from text associated with a specific index term, or derived algorithmically from the text of the document.

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

Spider
An automated process that presents documents to a data extraction or parsing engine by following links on web pages. (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 list
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.).

Stop words
Words that are deemed to have no value in an index. (See also Word exclusion)

Stopping distance
The point in a search query session where the user decides that time and effort spent in examining further results is not going to achieve additional relevant results.

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 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.

Thumbnail
An HTML rendition of a page from a document in response (often through a mouse roll-over) which provides the user with additional information about the potential relevance of the result.

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
The process of boosting index terms in specific areas of a document (for example the title) or on specific topics.

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
(See Stop list)

xAI
eXplainable AI is a set of machine learning techniques that produce more explainable models while maintaining a high level of learning performance and enable humans to understand, appropriately trust, and effectively use AI applications.