Supporting exploratory enterprise search with an AI/ML-based semantic tool

Lykke, Marianne

Published in:
Search Insights 2022

Publication date:
2022

Document Version
Også kaldet Forlagets PDF

Search Insights 2022

The Search Network

April 2022

Contents

Editorial services provided by Val Skelton (val.skelton@blythespark.co.uk).

Design & Production by Simon Flegg - Hub Graphics Ltd (www.hubgraphics.co.uk)

Introduction

The Search Network (TSN) is a community of practitioner expertise. We share a common passion for search, whether it be enterprise, e-commerce, intranet, web site or professional.

The Network was established in October 2017 by a group of eight search implementation specialists working in Europe and North America. There are now twelve members crossing the world from Singapore to San Francisco, meeting every month to exchange news and views on 'search'.

Search Insights 2022 is our fifth 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 the projects we have been working on, and make this knowledge open to the search community world-wide.

That is why there is no charge for this unique report, and why it carries no sponsorship. As in previous years we invite guest authors to contribute in areas where they have specific expertise. This year our guest authors are Professor Marianne Lykke, Patrick Lambe and Tim Allison. Since we commissioned Tim to contribute we are delighted that he has accepted our invitation to join the Network.

There is considerable diversity in the subjects in this edition. Some contributions look in detail at the technology of search, with the objective of helping you read between the lines of the marketing hype of software vendors. Others offer advice on implementation management and achieving project success. The list of search software vendors continues to grow, and there are now at least 75 companies we are aware of. We do not undertake work for any of these vendors so you can be assured that our advice is independent, neutral and informed from experience.

As well as this PDF version we have started to make individual chapters available on the TSN website. Based on user reaction we have included some of the contributions we have published in the past and taken the opportunity to revise them.

We hope that you find something of interest to you and your colleagues, and we hope we have the opportunity of working with you at some time in the future.

The Search Network

Tim Allison, Rhapsode Consulting LLC (USA)
Tim has been working in content/metadata extraction (and evaluation), advanced search and relevance tuning for 20 years. Tim is the founder of Rhapsode Consulting LLC, and currently works as a data scientist at NASA's Jet Propulsion Laboratory, California Institute of Technology. Tim is a member of the Apache Software Foundation (ASF), the chair/VP of Apache Tika, and a committer on Apache OpenNLP (2020), Apache Lucene/Solr (2018), Apache PDFBox (2016) and Apache POI (2013). Tim holds a Ph.D. in Classical Studies, and in a former life, he was a professor of Latin and Greek. Follow Tim on Twitter @_tallison.

Charlie Hull, OpenSource Connections **(USA, UK and EU)**
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 Haystack conference series 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.

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. She is the editor of "Taxonomies: Practical Approaches to Developing and Managing Vocabularies for Digital Information", published in 2022 by Facet Publishing. 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 out 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 company developing Datafari, its 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 teaches entrepreneurship at MsC students of the European Institute for Innovation and Technology (EIT). Prior to that, he spent ten years at SAP in the research department.

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. He is the author of Enterprise Search (O'Reilly Media, 2015) and is the Editor of Informer, the newsletter of the Information Retrieval Specialist Group of the British Computer Society. He has been a Visiting Professor at the Information School, University of Sheffield since 2002. Follow Martin on Twitter @IntranetFocus.

Guest authors

Patrick Lambe, Straits Knowledge (Singapore)
Patrick Lambe is a globally recognised knowledge management practitioner, and an expert in bringing KM principles to the discipline of knowledge organisation. Patrick was originally trained in Information and Library Science and arrived in KM via a second career in training and development. He is the author of "Organising Knowledge: Taxonomies, Knowledge and Organisation Effectiveness" (Chandos Publishing, 2007),

and co-author with Nick Milton of the award-winning "The Knowledge Manager's Handbook" (Kogan Page, 2019). Patrick is Visiting Professor in the KIM PhD programme at Bangkok University, former President of ISKO Singapore, and his next book, "Principles of Knowledge Auditing: Foundations of Knowledge Management Implementation" will be published by the MIT Press in early 2023. Patrick spends his time between Singapore and Dublin, Ireland.

Marianne Lykke, Aalborg University, Denmark

Marianne Lykke, Aalborg University, Department of Communication and Psychology is professor and leader of the research group e-Learning Lab. Her research interests include enterprise search, workplace knowledge sharing information architecture, and user studies. She acts as consultant for Danish companies and government agencies in the field of enterprise search information architecture. Follow Marianne on her research profile https://vbn.aau.dk/da/persons/122052.

Considerations for text and metadata extraction for search

Tim Allison, NASA's Jet Propulsion Laboratory, California Institute of Technology

Disclaimer

The research was carried out at the NASA (National Aeronautics and Space Administration) Jet Propulsion Laboratory, California Institute of Technology under a contract with the Defense Advanced Research Projects Agency (DARPA) SafeDocs program. Copyright 2022 California Institute of Technology©. U.S. Government sponsorship acknowledged.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

The following represents the viewpoints of the author and does not represent any of the funding agencies or reviewers.

In his article in Search Insights 2018, Martin White presented Udo Kruschwitz and Charlie Hull's schematic of a search system, as shown in Figure 1. White observed that search follows the "weakest link in the chain" model: "Poor performance in any of these major elements cannot be made up through superior performance in others".

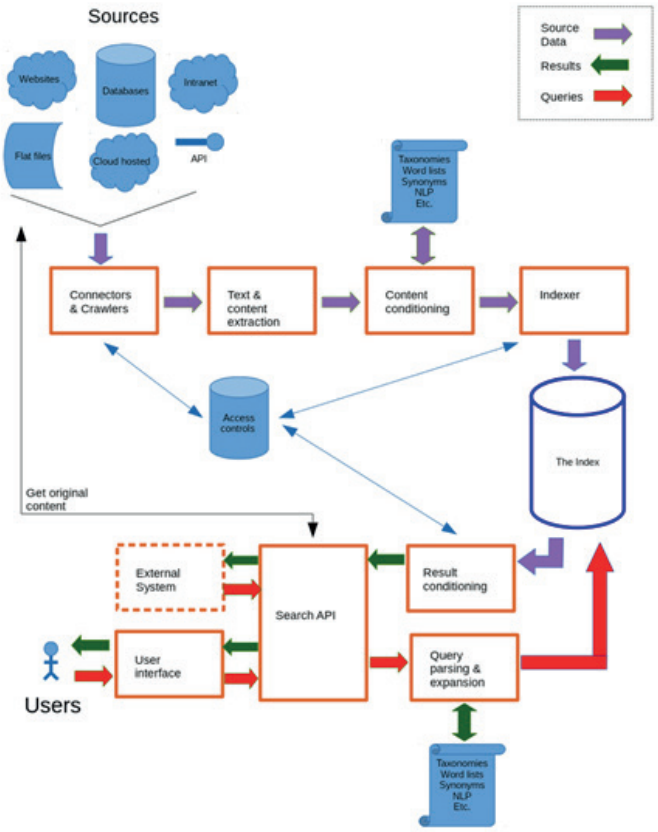


Figure 1: Kruschwitz and Hull's Schematic of a Typical Search System, originally published in "Searching the Enterprise", Foundations and Trends® in Information Retrieval: Vol. 11: No. 1.

In the very next figure in his article, "text and content extraction" vanish (Figure 2). The steps proceed straight from crawling to indexing.

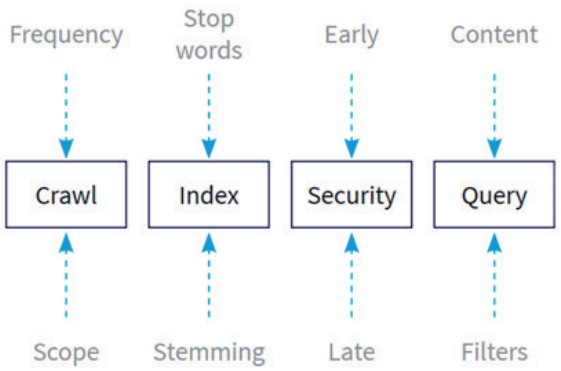


Figure 2: Martin White, "The technology of search." Search Insights 2018.

In this article, I'm grateful to have the opportunity to focus on text and metadata extraction as one critical step in the chain required for a successful search engine. As with the other components that White describes eloquently, when individual components fail or are not configured correctly for the data and/or user needs, search fails. However, unlike in other engineering endeavours, such as physical infrastructure, when search fails, users and developers rarely know that something went wrong. When a bridge collapses, people know it. In many cases, when a search system fails, it may fail silently – users will not be able to retrieve documents that they should have retrieved. Unless they know their document set very well, they may never know what they missed.

To bring the focus back to text and metadata extraction, extracting text and metadata from binary documents (Microsoft Office or Portable Document Format (PDF) files, for example) is non-trivial, and search developers often do not have the resources or tools to attend to this component of the chain sufficiently. In the following, I describe some of the things that can go wrong and offer mitigations where possible. I start from the easily observable to the more subtle and potentially more damaging risks, and I end with a special focus on some of the challenges of text extraction with PDFs.

Catastrophic failures

Parsing files is dangerous. In her LangSec 2021 keynote, Kathleen Fisher noted that the first and third most common Common Weakness Enumeration (CWE) in 2020 involved failures to validate input. She further noted that more than 1000 parser bugs have been found in Mozilla products related to PDF, ZIP and PNG. In an article for CMSWire, Martin White observes that "a rogue document might bring the (indexing) process to a halt."

When processing files at scale, catastrophic errors will happen, and content extraction pipelines must be hardened sufficiently to handle these failures without coming to a halt.

Catastrophic failures may include segmentation faults/crashes, out of memory errors, infinite loops, slow-building memory leaks. Mitigations and robust testing against this family of problems need to be built into the text and metadata extraction component. Carefully crafted files may also lead to remote code execution or other security compromises. Process-level isolation and containerisation are two techniques that can help limit serious problems.

Thrown exceptions

Parsers may fail to parse a document but not bring the whole process to a halt. For example, parsers may throw exceptions and exit gracefully. It is important to track these exceptions and identify the ones that are causing the most problems for your document sets. Parse exceptions are not randomly distributed, and if your data source is fairly homogeneous and contains files created by the same software, there is a chance that they'll contain similar features that cause problems for your parsers. A parse exception may be infrequent in one set of documents, but proliferate in others – "a small problem for you could be a big problem for me."

Of course, exceptions are also often thrown when documents are encrypted or otherwise protected against text extraction.

More subtle losses of information

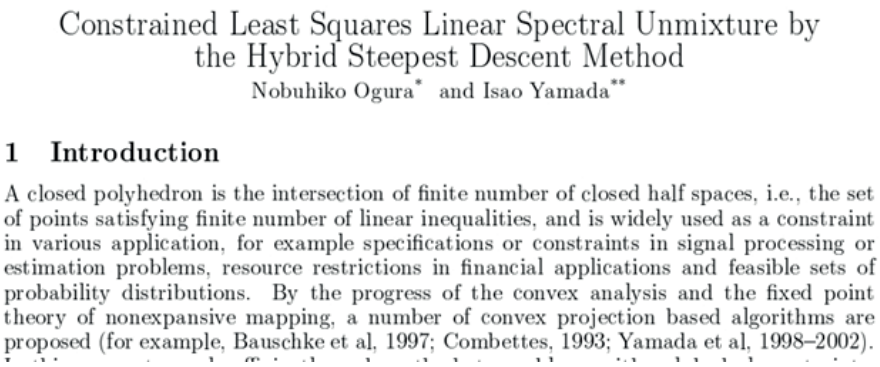
With the exception of security compromises, the problems mentioned above are all fairly obvious when they happen. The text and metadata extraction component must be built to limit the damage for the overall indexing system.

The more subtle and perhaps more dangerous problems are when a parser quietly fails to extract information correctly from a document. This may include not parsing embedded documents or not extracting metadata correctly.

Text may be skipped (partially or wholly), text may be extracted out of order (so called word soup) or otherwise corrupted. In these cases, information may not be searchable, and, as mentioned above, users may never know that there's a problem.

Without running multiple parsers for each document and comparing the output, there is no robust way to determine if a parser is missing text. However, there are some imperfect methods for indicating when a parser may have extracted corrupted text, a so-called "junk detector." Ashok Popat's A panlingual anomalous text detector lays out some methods for this kind of detection. The Apache Tika project's tika-eval module offers a calculation of the "out of vocabulary" statistic on extracted text. If extracted text does not contain a reasonable amount of dictionary words for a given language, there is a chance that the text has been corrupted.

To give a concrete example with a specific PDF file, let's consider 2002_Ogura_1_web.pdf. To a sighted human, this is easily readable (Figure 3).



Constrained Least Squares Linear Spectral Unmixture by the Hybrid Steepest Descent Method

Nobuhiko Ogura* and Isao Yamada**

1 Introduction

A closed polyhedron is the intersection of finite number of closed half spaces, i.e., the set of points satisfying finite number of linear inequalities, and is widely used as a constraint in various application, for example specifications or constraints in signal processing or estimation problems, resource restrictions in financial applications and feasible sets of probability distributions. By the progress of the convex analysis and the fixed point theory of nonexpansive mapping, a number of convex projection based algorithms are proposed (for example, Bauschke et al, 1997; Combettes, 1993; Yamada et al, 1998–2002).

Figure 3: Ogura and Yamada "Constrained Least Squares Linear Unmixture by the Hybrid Steepest Descent Method".

However, because of a flaw in the way that the PDF was constructed – it has no Unicode mappings – the text extracted is completely corrupt.

```
      !"#%$&  (')  *,+-).' / 0 1,23 *. 457698:;;<>=75?&@78;ACB
D(B7E;FHGJICBK5MLNBKOPBKF;B DJD Q R
S.TVU9WNXMY[Z\T]^W_S `badc 5KICedFgfh5 cji
:;edF;A^5KEk<>Imln:;e[<>EnloedACICe a lo<p57Eg5Kqsr;E;<jloe[E
8;O 6hedA5Kq adc 57ItedFk:;B c qsICf;B a
```

Figure 4: Ogura and Yamada – Text Extracted by pdftotext.

If a search system has no junk detector, this content will be indexed as is. This means that users will never find this document based on its content, and the search index will be filled with junk. If a search system does have a junk detector, this file will at the very least be logged as a problematic file and, perhaps, not indexed. The system may be designed to run Optical Character Recognition (OCR) on the file if there is a high out of vocabulary rate. On this particular file, Tesseract, an open-source OCR engine, extracts reasonable text (Figure 5).

```
Constrained Least Squares Linear Spectral Unmixture by the Hybrid
Steepest Descent Method

Nobuhiko Ogura' and Isao Yamada"
1 Introduction
A closed polyhedron is the intersection of finite number of closed
half spaces, i.e., the setof points satisfying finite number of
lincar incqualitics, and is widely used as a constraint in various
application, for example specifications or constraints in signal
processing or estimation problems, resource restrictions in
financial applications and feasible sets of
```

Figure 5: Ogura and Yamada – Text from Tesseract-OCR.

PDF-specific challenges

The type of corrupt text shown above may happen with other file formats, especially text-based files, including HTML, where the character encoding may be misidentified. However, given the ubiquity of PDFs, and their special challenges for text extraction, we detail some points to consider for processing of PDFs. For a more detailed treatment, see: Brief Overview of the Portable Document Format (PDF) and Some Challenges for Text Extraction.

Extracting electronic text from PDFs is a notorious challenge for those who work closely with the file format. In 2006, Michael Kay wrote the following on the challenges of extracting text and structure from PDFs (How we can convert pdf data into xml?):

> Converting PDF to XML is a bit like converting hamburgers into cows. You may be best off printing it and then scanning the result through a decent OCR package.

Figure 6: Michael Kay on Converting PDF to XML in 2006.

PDFs may contain any combination of images, vector graphics and electronic text; sometimes all three appear on a single page. If the images or vector graphics contain text, OCR is the option for making these searchable. Electronic text is not necessarily stored in logical reading order, but is rather stored to help the PDF reader render the text on the page. Because of this, some PDFs do not store spaces, but leave it to the

rendering software or the text extraction software to calculate where spaces should be inserted – this means that there may not be correct word breaks, for example.

Further, PDFs may not contain any notion of natural reading order or structural components such as headers or footers or tables. Because of this, extracted text may include header and footer text injected into the middle of a logical paragraph that was split over a page break. In Figure 7, we show a footer in 582116main_GRAIL_launch_press_kit.pdf and the text that was extracted from that file. For applications that expect logical sentences or paragraphs (such as natural language processing and machine translation), this kind of word soup can be extremely problematic.

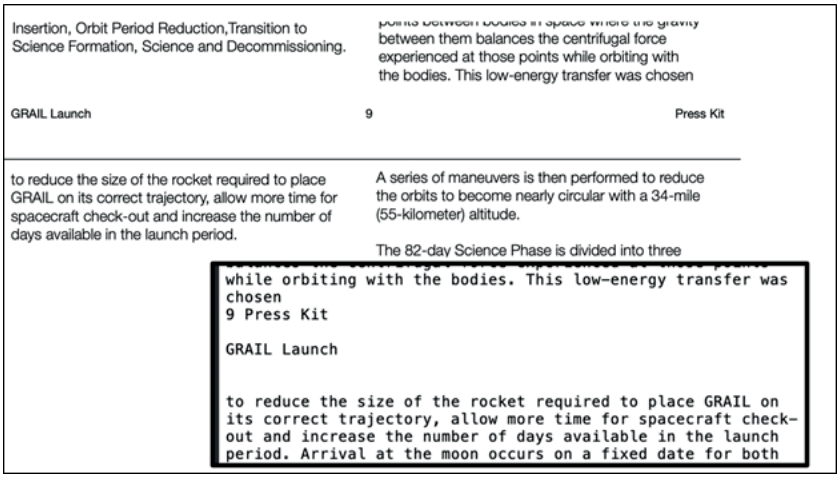


Figure 7: Example of Footer Text.

Many vendors, including Adobe (Adobe Launches PDF Extraction Generation APIs), have applied machine learning to reconstruct structural information from PDFs, and these techniques may help if your document sets are fairly homogeneous.

Another solution to this challenge is to generate the PDFs so that they are compliant with the PDF/UA (Universal Accessibility) standard, which was developed to allow encoding of reading order and structural components. Major companies, including Google, Apple and Microsoft now make it possible to generate PDFs that approach PDF/UA compliance.

Conclusion

Text and metadata extraction make up a critical component of search systems. File parsing is inherently dangerous, and pipelines must be engineered to be resilient against catastrophic problems. Problems with text extraction may be invisible or difficult to detect; some users may never know what they've missed because of failures at the text extraction phase. There are methods to identify potentially corrupt text, and there are workarounds once potentially corrupt files have been identified, including not indexing the document or, if possible, running OCR on the document. Given the scope of this article, we have necessarily skimmed over many of the details. Nevertheless, we hope to have communicated the importance of this step in the chain of components for a successful search system.

Skills for effective relevance engineering

Charlie Hull, OpenSource Connections

Search engine projects have historically been viewed as the responsibility of the IT department, covering installation, configuration, content ingestion and operations. The user interface may be developed by engineers themselves or with the participation of UX specialists. What this often leads to is a search engine that is performant and reliable but not necessarily accurate and in many cases is mis-aligned with business priorities.

At OpenSource Connections (OSC) we have long believed that without an effective search team, with members drawn from all areas of the business, you can only solve part of the problem. Your search solution is there to address a business case after all – for example to sell products, provide the correct information to users, save them time – but this business case is not always clearly presented to engineers. Search engines also depend very much on good quality content. No matter how clever the technology, this is very much a case of garbage in, garbage out. New, exciting features (nowadays often incorrectly presented as AI) will not help when metadata is inconsistent and data modelling incomplete.

So when you're considering your next search project, what sort of skills should you look for in your search team? Let's consider the roles in a 'perfect' search team. (In smaller organisations the same people may hold several of these roles, and it is rare in our experience to find an organisation that has all the roles covered.) This list is drawn from OSC's 'Think Like A Relevance Engineer' training materials.

- **Stakeholder** - responsible for aligning search improvements with financial and corporate benefit
- **Product Owner** - responsible for ensuring search improvements meet the information needs of the customer
- **Project Manager** - responsible for planning and prioritising changes that are translated to features from the customer information needs
- **Product Developer** - responsible for design and UX implementation in the product
- **Content Owner** – responsible for defining the content set for the product, and coordinating development teams to arrange content access and transport
- **Metadata Owner** – responsible for defining and managing any metadata assets that are used to improve search, including synonyms, lemmatisation files, spelling dictionaries, word-wheels, etc.
- **Architect** - responsible for integration strategy and planning of technical changes across the system for cross cutting concerns and big-picture technology fit
- **Search Relevance Strategist** - responsible for solution strategy and planning of technical changes across the system related to search improvements
- **Search Relevance Engineer** – responsible for search engine tuning and delivering associated measurements and experiments
- **Software Engineer** - responsible for solution delivery and detail-oriented implementation of functionality and features related to search improvements
- **Data Analyst/Scientist** – responsible for analytical data access and transport, identifying customer trends and engagement signals, and coordinating judgement and rating data acquisition

You may already have some of these roles filled in your search team, and some can be generic across many projects – Product Owner for example. However, some of these roles need specific, specialist skills.

A **Search Relevance Strategist** is someone with long experience of information retrieval, search engine technology and implementation. They know what is cutting edge technology, but also what basic functionality should be built first. They know how to measure search quality effectively and how to design processes to make this happen. They probably don't write production code, but they can guide and mentor others on the team who do. They have good communication skills and can inspire others to improve search quality.

A **Search Relevance Engineer** has practical, up-to-date knowledge of the search engine you are using, is aware of its features (and drawbacks) and how to implement them to solve relevance challenges. They are trained in information retrieval fundamentals, know how to model source data for effective search and how to format search queries correctly. They can be a highly effective member of your development team, using their experience to build features, right first time.

A **Data Analyst/Scientist** working on a search project needs to know what can be measured, what is important to measure and what conclusions can be drawn from the data. They can help you create meaningful metrics and visualisations so the whole search team can see the impact of a potential change and identify potential risks and benefits. Search quality is not always easy to measure, and you may not have (or be able to gather) a full picture of how your users interact with a search application, so a pragmatic approach is best. A search data analyst/scientist will have a good grasp of the various relevance measurement tools that have appeared over the last few years, many of them open source software.

Content Owners and **Metadata Owners** are usually subject matter experts (SMEs). SMEs know your content intimately – in e-commerce search they know what you sell (and importantly what you don't), what your competitors sell and what is the 'right' answer to a search query. They're familiar with part numbers, content areas, what is up to date and what is a little stale. They're a helpful librarian who knows which shelf holds that obscure book you can half remember; a gardening expert who knows which fertiliser to use on your roses; a legal taxonomist or a medic aware of the difference between a cardiologist and a cardiothoracic surgeon. They are aware of the structure of your content – which fields are important to search and which are additional context. In your search team these SMEs can explain to engineers why a result is good, or bad for a particular query and they can be essential parts of any search engine test framework, giving expert ratings (but not always agreeing with their colleagues on these).

So now we know how to build the perfect search team, how do we make sure each member has the appropriate skills? As many have discovered when trying to recruit staff for their search project, these skills are relatively rare, and experts can command a high price. OSC's advice is to focus on empowering your search team for success, supporting them to develop their skills so eventually they can fully 'own' the search solution. There are various ways to do this:

- Expert training. OSC and other organisations provide training in search engine basics relevance engineering and some more advanced topics such as Learning to Rank and Natural Language Processing. The quality of this training can vary, especially at the beginner level, and we recommend you consider training delivered in a practical fashion with exercises and workshops.
- Partnering and mentoring. If you work with external partners, think about how they can help mentor your team and teach them the skills they will need while they work

on your project. It can be a bad strategy to outsource your search entirely as it can lead to over-dependence.

- Read the literature. There are many books and useful blogs on information retrieval search engine fundamentals, relevance engineering and even advanced topics such as Deep Learning for Search. Some good starting points are available on Martin White's website.
- Attend events. There is a range of events where search and relevance topics are discussed, ranging from large commercial conferences such as Lucidworks' Activate and Elastic's Elasticon, academic events such as ECIR and SIGIR , smaller and more community-driven events such as OSC's Haystack, the British Computer Society's Search Solutions and Berlin Buzzwords. In many larger cities there is a regular Search Meetup which is usually free to attend. Encourage and fund your team to attend these events, meet others facing the same challenges and importantly, participate by presenting or even offering to host a Meetup.
- Interact online. There are mailing lists and forums for particular search engines such as Apache Solr and Elasticsearch and more general forums such as OSC's Relevance Slack. Since the search community is widely distributed these can be a good way to keep in touch with others outside of physical events.

In conclusion, an effective search team will be drawn from areas across the business and will require members to have a wide range of skills. The specialist skills required to improve search quality are rare and we recommend that you support your own staff in gaining these, using external partners where necessary, but being aware that you should aim at eventually owning your search. Community participation is vital and helps both with skills development and also to publicise that your own organisation has committed to building effective search – which can help when recruiting and retaining staff, as well as promoting the technical excellence of your approach.

1 http://intranetfocus.com/search-resources-books-and-blogs/
2 https://activate-conf.com/
3 https://www.elastic.co/elasticon/
4 https://ecir2022.org/
5 https://sigir.org/
6 https://haystackconf.com/
7 https://www.bcs.org/membership-and-registrations/member-communities/information-retrieval-specialist-group/conferences-and-events/search-solutions/
8 https://berlinbuzzwords.de/
9 https://lucene.apache.org/solr/community.html
10 https://discuss.elastic.co/
11 relevancy.slack.com/

This is a revised version of an article that first appeared in Search Insights 2020.

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

Charlie Hull and 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 objec-

tives 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.opensourceconnections.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

This is a revised version of an article that first appeared in Search Insights 2021.

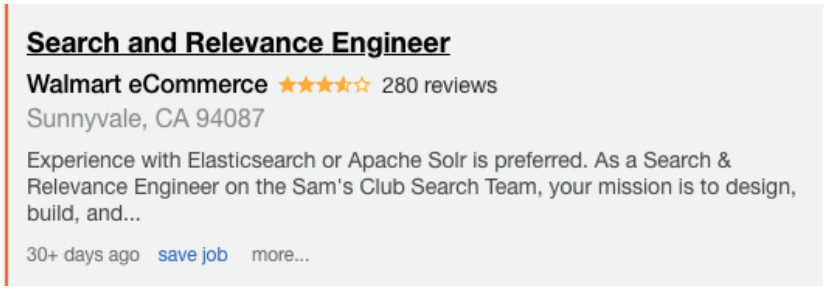
The rise of the relevance engineer

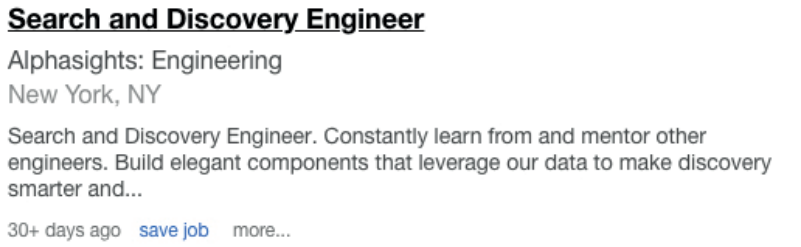
Charlie Hull, OpenSource Connections and Doug Turnbull, Shopify

Perhaps you've noticed a trend lately.

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

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





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

What is a relevance engineer?

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

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

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

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

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

Where did the 'relevance engineer' come from?

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

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

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

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

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

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

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

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

The relevance engineer's persistent problem: measurement

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

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

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

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

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

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

Always with relevance engineering: arriving at a good solution requires more than technical expertise, interpreting and acting on user feedback requires collaboration with product stakeholders. Understanding the context a user is operating within, for a spe-

cific application, that prompts them to type 'december projections' takes a tremendous amount of skill in the domain.

How can you grow your relevance engineering capabilities?

Despite the success of *Haystack*, there is still a huge need for search and relevance expertise. Demand outstrips supply, with far more jobs available than applicants to fill them.

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

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

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

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

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

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

This is a revised version of an article that first appeared in Search Insights 2019.

Project budget

Miles Kehoe, New Idea Engineering and Charlie Hull OpenSource Connections

A structure for a search budget

Enterprise search vendors are very reluctant to give any indication of the likely budget needed to implement their software. There are a few exceptions, notably Mindbreeze (which is a search appliance) and dtSearch, which has always had a very transparent pricing structure. What is virtually certain is that every vendor will have a slightly different way of presenting the costs associated with implementation so direct comparison is going to be a challenge.

If open source software is chosen, then there is no vendor. This removes some cost elements and you might reasonably expect that the overall cost will be lower. However, much more software development may be required, raising the cost of professional services and also raising internal costs. Beware of the misconception that open source software is 'free' – it is 'freely available' but adopting it is not cost-free.

The main cost elements are

- Vendor software licenses
- Vendor professional fees
- Vendor maintenance and support fees
- Third-party software license fees
- Professional services from an implementation partner
- Internal costs for hardware, software and the support team
- Proof of concept/validation charges

Vendor software licenses

Vendor software licenses are almost always volume related. The volume elements can be the number of documents/content items, the number of users or file storage capacity. In the worst case each entry in a database and/or each individual email may be represented as a separate content item in the search engine, so these numbers can be very large. To arrive at a figure for the number of documents requires a content audit because the IT department may well know the amount of storage but not the number of files. The format of the files will also be important. The number probably needs to be 'correct' to the nearest million. Arriving at a static current number is going to take time but extrapolating over the next three–five years may be more difficult. Many of the files will be versions or copies but these all count as a content item when it comes to indexing.

Another factor is how far back the search application should go into archive material, where file formats may not be standard Microsoft Office and metadata tagging is likely to be partial or non-existent.

The number of users may seem to be quite straightforward until a vendor asks you to categorise types of user: broadly speaking users who are undertaking a simple Google-like search and those who are going to spend a considerable amount of time at each search session. Another issue is that many companies may have a substantial number of employees in manufacturing or out on site as service and support staff. Their use of search may be very limited, but they will still expect access when the need arises. Any such categorisation is very difficult to define, agree and monitor.

The time period for licenses will also be open to discussion. Many vendors offer a perpetual license (which has no expiry date) or an enterprise license (which requires little or no clarity on the number of content items or the number of users). These and other variants require the customer to have considerable clarity about the future which is close to crystal-ball gazing. Any company that gets to the stage of negotiating a contract without a well-prepared three-to-five-year search strategy is either going to pay too much or buy too little.

Vendor professional fees

As with any enterprise software application the professional fees will almost certainly be substantially more than the software costs. Installing and implementing search software is a highly skilled business and the engineers working for the vendor will be paid well because the last thing that the vendor wants is for a skilled engineer to leave. Replacing them can be time-consuming and there is a substantial amount of training to be given. Gaining clarity on the roles that the professional services team will play is very important. Some elements of the service provision may be fixed fee, but it is more normal to be charged on a time and materials basis. There could also be a substantial travel and subsistence costs if engineers need to be flown in from the USA to Europe, or any other long-haul journey. These costs (and the same for implementation partner costs) will inevitably be front loaded.

Vendor maintenance and support fees

Maintenance and support fees will often be presented as a series of packages with an increasing level of support. Often there is a lot of attention paid to '24/7' support with response times of a few hours. This usually refers to being able to register a ticket with a human being and not the time taken to solve the problem. Few vendors are willing to commit to a resolution timescale. There is no reason why a prospective purchaser cannot request a customised service package which integrates well with the level of service that can be provided by a well-trained IT team.

Third-party license fees

Search applications are very modular, and there could be a requirement for specific applications to meet a customer requirement which is best met by using a product from a third-party vendor. Examples would include high-end analytics visualisation, specialist connectors and linguistic products for non-European languages.

Professional services from an implementation partner

Relatively few search software vendors will undertake all elements of an implementation. Their business is in developing and selling software, not professional services. Almost certainly there will be a requirement for additional support, perhaps because the implementation is global, but the vendor operates only in the USA or Europe. This support could come from an existing implementation partner or more likely would come from what is a very small number of specialist search implementation companies that have partnership agreements with one or more vendors. As is the case with vendor professional services staff, partner engineers are also in short supply so day rates of €2500 are quite usual.

The involvement of an implementation partner raises the issue as to whether there are separate contracts with the vendor and the partner, or a contract with just the partner, who then supplies and implements the software.

Again, these professional services fees will almost certainly be on a time and materials basis, and that makes budget forecasting by the customer very difficult.

Internal costs for hardware, software and the support team

Enterprise search applications are development platforms and not products, so a development + production environment will be the norm. Network bandwidth can be an issue – users expect response times of no more than 500 milliseconds wherever they are on planet Earth, a requirement that few other enterprise applications will need to match outside of e-commerce applications. There could well need to be additional hardware and associated operating licenses to be procured and additional network capacity bought in for core global operations in (as a good example) China.

To a significant extent the quality of search performance is a function of the investment in search support staff. The team costs (which exclude IT staff!) may not need to be taken into account but if they are, the percentage addition on the cost could be considerable. For companies of between 5000 and 10,000 search users (i.e. excluding manufacturing) you can probably get away with three full-time staff on the search team, even if not co-located. Above 10,000 users then you may need at least four. An experienced enterprise search manager can command a salary of at least €120k, and so a team of four could be at the €500k salary level, or €1m if overhead allocations are taken into account. Recruitment costs may also need to be budgeted for.

Proof of concept/validation costs

Some vendors will make a charge for the usually two-month long proof of concept phase but may also set this against the costs for the implementation. Others will not make a charge or will ask for a fixed price fee to cover administrative and related costs.

How much?

Because of the number of elements in the cost structure it is very difficult to give a good answer to this question. If you have 10,000 users (or more) and perhaps 30 million files to index, then the external costs from the vendor and implementation partner could be of the order of €2-3 million over a three-year period. However perhaps 60% or more will be professional services costs, and the scales of these will not be obvious until well into the contract negotiation. As with all software and services contracts everything is open to negotiation. Factors that could bring down the price include offering to act as a reference site, investing in a wider range of skills in the project team and offering the vendor an opportunity to gain a blue-chip client in a market sector that is core to their business development.

How long?

Contract negotiation will almost certainly bring up the issue of the respective scopes of the vendor and the implementor, as well as subsidiary contracts with third-party software suppliers. It is advisable to develop a draft structure to the contracts at an early stage, often referred to as a Heads of Agreement. Even then expect the negotiations to take at least two months from the Heads of Agreement to the Signed Contract. There is an immense amount of detail to get through and your Procurement Department will probably never have worked with the vendor before, so there are no established lines of contact and trust.

This is a revised version of an article that first appeared in Search Insights 2018.

Enhancing search with context mapping

Patrick Lambe, Straits Knowledge

At its most crude, search looks for word matches between the text of the query and the text in the content being searched. It pulls together results on the assumption that word matches indicate the documents are about similar matters to the query. This is crude, because of the vagaries of language. The same word can refer to very different concepts (sometimes contradictory) and the same concept can be expressed using different words or spellings. How will an untuned search engine know that "Kyiv" refers to the same city as "Kiev"? Does "sanctions" in this document refer to deterrents meant to prevent an action (the EU imposes sanctions on Russia), or does the term refer to an act of permission or approval (Putin sanctions an invasion)?

There are three components to a really strong search toolkit. Used in combination, they can dramatically enhance the user search experience:

- Advanced search tools and strategies
- Knowledge organisation systems
- Context mapping

Advanced search strategies and tools

Advanced search strategies and tools can mitigate the weakness of crude word-matching. Advanced search protocols such as Boolean expressions can force matches with terms expressed as phrases, or excluding certain term combinations, giving more precision to the search definition. Vector search and other tools based on text analytics can look for document similarity based on strings of text, and relative adjacency between terms (see the article by Charlie Hull and Eric Pugh page 47). Search auto-suggest can guide users towards known high-value search expressions using curated lists (see Avi Rapaport chapter in Search Insights 2021).

Knowledge organisation systems

Knowledge organisation systems (KOS) – which include taxonomies, ontologies, knowledge graphs and thesauri – bring known relationships in the world to a search index to enrich it and benefit the person making search queries. Even more powerful benefits accrue when a KOS is effectively used in content tagging. Good metadata can foreground the most relevant search results, if the KOS accurately reflects the working environment that it represents. The benefits of a taxonomy, for example, include (for more, see Helen Lippell's article, page 30):

- **Hierarchical structures** – such as parent-child, whole-part, class-member, process-step, topic-subtopic – allow a very general search to be narrowed down using refiners, or a very narrow search to be broadened, because the hierarchy can guide the searcher between a general concept (e.g. Ukraine) and more specific ones (e.g. Kyiv, Kharkhiv, Donetsk) even if those terms are not consistently present in the documents being gathered together. The same function allows documents about more specific concepts to be "rolled up" together so that they can be explored in any direction the hierarchy allows.
- **Synonyms, variant spellings, acronyms and variants in different languages** – can allow searchers to find content associated with a single taxonomy concept, even if they use very different search queries that are not present in the target documents, or if different documents use different terms, as long as they are mapped in the background taxonomy.

- **Associative relationships** – can be mapped between concepts in different parts of the taxonomy structure so that events (the Russian annexation of Crimea in 2014) can be associated with other events (the broader invasion of Ukraine in 2022), with people (Vladimir Putin, Volodymr Zelensky, Alexander Lukashenko), with relevant institutions (NATO, European Commission, Federal Assembly of Russia), and with relevant laws, regulations and treaties (Montreaux Convention, U.N. Charter).

Context mapping

There is a third, critical piece in the toolkit that can bring relevance and precision to search. This is a good understanding of the contexts of information use. Even within a military context, the meaning of "tank" can refer to different things depending on whether I am conducting a search within the context of logistics and supply lines (e.g. a large container for fuel), or armoured formations (e.g. an armoured fighting vehicle). A KOS with synonyms can provide one means of disambiguation, but if a system knows who the user is, and what the normal context of her work is, then it should be able to direct her immediately to the most relevant results based on that context, without any visible need for disambiguation.

This is exactly how Google Search works – in this case using geo-location as a contextual determinant. If I am located in Ireland and I type "1500 SGD", it will immediately give me the Euros equivalent, because it can accurately guess that's my interest. If I am in the USA and search for Paris, it may suggest Paris Hilton before Paris the city.

Understanding the contexts of key audiences is a large part of how we go about designing a KOS in the first place. We look at user contexts as well as the range and variety of content (and the language used within the content) when we are designing a KOS.

But even within a broad-based KOS that covers multiple functions, we may want to channel only the most relevant parts of the KOS to a relatively specialised functional area, and we may want to tune the search interactions to that functional area, so as to remove the "noise" from other areas and sharpen the focus of search.

How then do we get a good understanding of context, and how can we exploit it in search? Here are four different approaches.

- Use case scenarios and content modelling
- User journey mapping
- Business process mapping
- Knowledge resource mapping

Use case scenarios and content modelling provide a very targeted method. We select a specific kind of user and a key task or function that they perform. Perhaps we are looking at a central bank and their inspection of financial institutions. We can describe the steps in inspections, and the key sources of information they use, and the documents they produce. We can use this to model the information architecture, the content types they will use, and the key metadata they need, including KOS-related metadata.

All of this information can be used to collect and map synonyms, define search refiners and search result sources, compile search auto-suggest lists, and infer related search suggestions using associative relationships.

If we structure the information environment to follow the steps in an inspection pro-

cess, this can also be used to automatically add relevant KOS metadata to documents that are contributed at each stage of the inspection process. The system knows who you are and where you are in the process, so in principle it should (i) be able to add relevant metadata to new documents and (ii) tune the search results and search functionality to that functional area.

The limitations of this approach are that it is labour intensive to design, set up, and maintain. The benefits are that it provides a very high quality of metadata for search and discovery purposes even with largely unstructured documents. If the functional area being supported is sufficiently high impact (in terms of value, risk or scale) then the effort will be worth it.

Business process mapping is a slightly more generalised form of use case mapping and content modelling. It follows and maps major business processes, and their interactions, and this can be used as a means to identify and map the major document and information content associated with these processes and their sub processes.

These maps can be used to tune both content types and the search functionalities to the specific working contexts described in the business process maps. Users can be guided through the process, seeing only what is relevant at each stage, and having easy ways of navigating to next or prior steps.

Similar to use case mapping and content modelling, business process maps can be used to design a supporting KOS and information architecture, as well as automated metadata collection mechanisms, depending on where in the process you are contributing documents. This approach is especially useful for users who have to work within business processes, and it is designed to help them perform their tasks and roles within it.

User journey mapping is similar in principle to business process mapping, except that it comes into play when our target users have to navigate *between* back-end business processes. For example, citizens who consume government services where different agencies may be involved and different processes intersect (e.g. marriage, citizenship, immigration, social security, health, law enforcement websites). In this case the maps model the key journeys that users undertake, and similarly to business process mapping, will map and harmonise the vocabularies, content types, and transactions involved. On corporate intranets, a similar approach can be taken for employee services – e.g. where processes and policies regarding employee benefits, insurance, payments, annual leave and performance management may intersect, but are managed by different corporate functions.

Knowledge resource mapping has affinities with both business process mapping and user journey mapping, in the sense that it generally tracks key activities, and then the information or knowledge required and produced within these activities.

Knowledge maps are typically created within knowledge audit processes as part of a broader knowledge management initiative. They provide more depth and detail about the nature of the knowledge that is being used within the groups being examined. So for example, they can track knowledge flows between key parties, and they can track where tacit knowledge in people is being deployed, and not just the explicit knowledge that is codified within structured or unstructured content.

These maps can be used not just to map content to very structured steps in a process or a journey, but also to understand motivations and causes for less structured and more exploratory search and discovery interactions. Knowledge maps are richer than process or journey maps, and make it much easier to map associations between activities, key documents, people, institutions, and events. For this reason, they are very powerful context-building aids in KOS design, as well as in search functionality design.

Knowledge mapping is the most labour intensive approach of the four outlined here, so it is most often used in more complex information environments, when designing an enterprise KOS, information architecture and search architecture – i.e. where the organisation needs both breadth (to pull together related content across many information stovepipes) and depth (to understand and model specialised functions using narrower ranges of content and vocabulary). In this way they can support more targeted search and discovery designs and yet still work effectively and consistently within a single architecture.

Further benefits

Context mapping is not just useful for the initial design of search and discovery services (and their supporting KOS). Because context mapping provides clear and consistent characterisations of user needs and typical patterns of behaviour, these context maps can also be used to measure search effectiveness. They can be mined to develop measurable search scenarios (how quickly and successfully specific documents are retrieved by a cross section of users) and measurable discovery scenarios (what specific insights about the coverage of a collection can be gathered from a resource base by using search queries).

To sum up, context mapping is an extremely valuable approach for improving the “last mile” in search effectiveness, for both search and KOS design.

Mapping taxonomies – benefits, case studies and how to do it

Helen Lippell

Introduction

Any taxonomy project should be done from a foundation of wanting to develop common vocabulary and understanding across the organisation (and outside the organisation, for taxonomies being used for external-facing websites, apps, services etc). Often I get asked to develop a single taxonomy to "rule them all", to rip off the Lord of the Rings quote. Can work perfectly adequately for taxonomies that have a limited number of broad categories - an example of this might be a records management taxonomy that only needs to classify documents into an agreed number of types or purposes.

But increasingly it is becoming necessary to devise solutions for tagging that allow data from multiple sources to flow between systems, which means that different taxonomies (or other knowledge organisation models) have to be mapped. Different taxonomies are compared methodically, and each taxonomy term is mapped, if it can be, to a term in another taxonomy.

This is no mere intellectual exercise. This article will look at the benefits of mapping, the practical considerations of carrying out mapping and consider some real-life scenarios. In this article the terms data, content, document and information are used more or less interchangeably to mean 'stuff' that needs to exist in multiple business systems.

Three key use cases for taxonomy mapping

Improving search

A common reason for carrying out taxonomy mapping is for enhancing search quality. This can apply for public website searches and enterprise search. Inside organisations, federated search is an application that returns results from multiple sources, perhaps a knowledgebase, an intranet, content management systems, to name some possibilities. This unified search can use the taxonomy terms and their mappings to return more consistent and relevant results for search queries. Often, users may not know exactly in which system the most relevant results for their query are. Different systems may use different taxonomies because they were designed for different use cases, and developed at different times.

Taxonomy mapping was used in a search project for a conservation organisation. The organisation had a specialist team that was using scientific, industry-standard taxonomies when tagging content. The vocabularies covered detailed names and taxonomic hierarchies for flora and fauna. Other teams wanted to use simpler vocabularies that included common names and nicknames, as they did not always need to go to the level of granularity of, say, the three species of voles in the UK; it might be sufficient just to tag a blog post with 'voles'. (If granular detail was needed, then the teams would liaise to ensure accurate tagging.) Mapping the scientific names to the common names in the two vocabularies, searchers could access content from technical publications, as well as blogs written for a general audience.

Supporting interoperability

Interoperability is another rich area of potential for taxonomy mapping. This is about the flow of data between systems, in order to enable business processes or information exchange. An example might be the need for procurement, finance and supply chain

management systems to be able to share data about the contracts, money and products that keep a business running. The classifications and code lists connected to this data should be mapped across the different systems so that, for example, a supplier's relationship with the organisation can be tracked right from tendering for work, to when their invoices are paid, and through multiple contracts.

Interoperability can also be needed between organisations. I've delivered projects for a couple of international public sector organisations that operated across European or global borders, and worked closely with other bodies such as government departments, regulatory agencies, service providers and so on. The organisations needed to map taxonomies from both internal and external sources, containing quite technical language, in order to ensure documents were categorised correctly, and could be retrieved consistently even if stored in different systems (such as knowledgebases, intranets and records management systems). Some taxonomies also contained translated terms (see the next section for some considerations around mapping multilingual vocabularies). While the technical effort needed to do mapping in these organisations was substantial, the business aspects took even more work (see below).

Enriching and integrating data

The third use case is data integration, that is, the process by which data from multiple sources are brought together in one system as a unified view. This is often done in order to build a product or service. An example, explored more fully in the next section, is ingesting data feeds from an external source, and combining them with data created by the business, to produce a database that users can search to find what they want. This could be structured product information from manufacturers and wholesalers, that gets combined with content about the products for an ecommerce site.

Data enrichment using open or Linked Data comes under this use case too; Linked Data services (such as DBpedia, perhaps) can be mapped to the organisation's own data to create something new that offers more value to the user than just the Linked Data or proprietary data sets on their own. An example might be a news agency (or other information business) enriching their own data with the contents of publicly-available Linked Datasets.

I worked on a data integration project for an entertainment listings provider. The company was buying data feeds (say, cinema listings) from third-party specialist providers and then aggregating them into searchable products for users. The feeds could also be enriched by the company's own data and content, such as film reviews from their critics or automatically-generated restaurant recommendations from their own algorithms. The data, including categories, topics and code lists, had to be normalised and mapped to the company's one global taxonomy. For example, where genres in the external feeds and genres in the global taxonomy differed, they would be mapped wherever possible. Another interesting data modelling quandary was to map event times based on whether they should be displayed in a feed of events happening 'at the weekend'. When does the weekend start, and when does it finish? For some people, the weekend might start on Thursday evening, or immediately at the end of a working day on Friday, and the concept of a 'weekend' differs in other cultures too.

Finally, there is a lot of overlap in how data feeds describe entertainment venues, particularly in dynamic cities. There may be technical definitions of places like bars, restaurants, pubs, gastropubs and nightclubs (e.g. for planning purposes), but ordinary users will just have a rough idea of what kinds of experiences they will get in those venues. If the wrong kinds of venues are returned from a search, they will just search on another site, as there is a very crowded market for entertainment information.

Simple Knowledge Organization System (SKOS) mapping

It might look deceptively straightforward to map term A in one taxonomy to term B in another taxonomy, but the devil may still be in the details. A foundational principle of any controlled vocabulary is that they are needed because language is ambiguous. The same word can have multiple (even contradictory) meanings, and different words can have the same meaning. Beyond that, words can have contested meanings, culturally or contextually diverse meanings, or the boundaries between the everyday definitions of words can be blurry. The next section has good examples of the latter scenario.

To help with mapping, the SKOS (Simple Knowledge Organization System) data model, which can be used for modelling for all kinds of knowledge organisation systems, has five properties that can be used to specify the degree to which two terms in different taxonomies align. This goes beyond what can be expressed in traditional thesauri, which only allow RT, or Related Term. In the following examples, imagine Taxonomy A is a list of major world cities, while Taxonomy B covers French geography:

SKOS property	Term in Taxonomy A	Term in Taxonomy B	Notes
Exact match	Paris (France)	Paris	(Terms can be mapped as Exact even if their labels are different)
Close match	Paris (France)	Métropole du Grand Paris	Wider administrative area that also includes outer suburbs of Paris
Broad match	Paris (France)	Île-de-France	The region that Paris is part of
Narrow match	Paris (France)	Stade de France	A stadium in Paris
Related match	Paris (France)	Marseilles	French cities

An extra dimension to mapping is multilingual taxonomies. All the fuzziness of semantic ambiguity in one language can be amplified when trying to translate words between languages. There are many concepts that are used in English but that have no perfect translation from their original language, e.g. schadenfreude, hygge, chutzpah or kawaii. When non-English words are not used in English at all, it is difficult to map them to something in an English vocabulary.

Practical considerations

The technical aspect of taxonomy mapping is just one part of the work. It is important for taxonomists and others on the project team to build solid relationships with stakeholders, including the owners of other systems, the managers signing off on mapping work, or the developers who will be implementing the mappings. It needs understanding and cooperation on all sides, even more so if the mapping is part of a larger digital transformation programme. Processes for making the actual data flows happen need to be established, otherwise elegantly-mapped taxonomies will have been done in vain.

As with any taxonomy work, the questions of governance and maintenance are critical to long-term success. Taxonomies will change over time, therefore any mappings contingent on them will need to be regularly reviewed. Otherwise there is a risk of broken data and divergence between systems. Strategic reviews will be needed too as new opportunities may be available, for example, new channels to send content to, or new taxonomies/Linked Data to map to.

Conclusion

I've deliberately not gone into the role of ontologies in this article, because I wanted to focus on mapping vocabularies rather than classes of things. That said, there is a lot of potential in using an ontology across the organisation to bring out some of the benefits of using consistent definitions and labels for things the business cares about. (For the conservation organisation, an example might be 'birds', for the international body 'document'.)

Knowledge organisation models that are designed for individual use cases, but adaptable enough to link to other models, are better than trying to shoehorn solutions for all needs into a single taxonomy. I've seen a few taxonomies that looked like Frankenstein monsters because they had had bits added over the years, and as a result, were not serving the needs of the business or users. It's preferable to design taxonomies mindfully, and with regard to possibilities for mapping to other taxonomies in future. Mapping can be a delicate exercise in terms of both technology and business considerations, but it's clear that the pay-off for the effort can be more than worth it.

Good practice in taxonomy project management

Helen Lippell

Introduction

Over the years that I've been Programme Chair of Taxonomy Boot Camp London, I've noticed a healthy increase of interest in the business use of taxonomies, ontologies and knowledge graphs. Leading the development of the Programme means I encounter many examples of good practice and excellent real-world implementations. Taxonomy is well and truly breaking out of its roots in the library and information science worlds to find uses in all sorts of digital applications.

In the private sector, taxonomies are business assets, constructed to support products and services that are intended to stand out in their marketplace and, one assumes, earn revenue for the organisation. Taxonomists want information management to be recognised as a core business process, and as something of great value. If people can't find what they're looking for on your website or app, they will go somewhere else. Very few companies have totally unique and proprietary information or data to sell.

In cases where taxonomies are used to help users inside an organisation find and use information, these are often developed because an organisation wants to reduce inefficiency. Taxonomies in the public sector may be needed for either internal or external use, but again are built in order to achieve a specific outcome that benefits the business.

Here I explore some of the reasons why taxonomy projects fail and what taxonomists and others can do to ensure they don't. The issues and mitigations discussed are equally applicable to enterprise search and ontology/knowledge graph projects. The approaches described are designed to be useful whether someone manages a taxonomy as part of a wider role, is a dedicated in-house resource, or is a consultant.

The stages of a taxonomy project

There are many potential bumps in the road between a taxonomy project being conceived, and the taxonomy being a long-term success. It's rarely a smooth or quick process to get a taxonomy or search project approved in the first place. Once the project is completed, the work of embedding a taxonomy properly is usually part of a wider change management initiative. Metaphorically flinging a new taxonomy 'over the garden fence' and expecting users to adopt it from day one, without helping them to understand the business benefits, rarely ends well. In the end, without an ongoing plan to keep the taxonomy for its intended purpose (or purposes), then the investment of time and money will have been squandered.

Business approval

Starting at the beginning, there are many ways that taxonomists can bring colleagues along with them on the 'journey'. It comes down to listening, educating and advocating as much as possible within the organisation. Every conversation is an opportunity to sell the benefits of taxonomies, not just at the senior management level from where a project sponsor might emerge, but also at the peer level, where colleagues might be wary of how a taxonomy could change their current ways of working.

(One thing that can happen during the project is that the sponsor changes, causing uncertainty and even risking the completion of the project. Business priorities may change as a result. Sponsor changes are outside a taxonomist's control. Other than

continuing to cultivate understanding among a range of senior managers, there's not much we can do with this one.)

Finishing the initial phase of development

Implementing a taxonomy, whether on its own or as part of a larger technology project, is a reason for celebration. The hard work of gathering requirements, understanding user needs, understanding the domain and analysing the content or data has been done. This knowledge has been translated into a live, working taxonomy. Everything is cool now, right? Not necessarily.

The most common failure scenario is that once the taxonomy project team has completed its mission, it is disbanded without a commitment to ongoing maintenance of the taxonomy. The project checklist item 'do the taxonomy' is ticked off and not enough consideration is given to day-to-day operation and the maintenance of quality.

Moving into business-as-usual

There is a common misconception that a taxonomy is just 'finished', yet most domains can and do change over time. Terminology moves on, new entities and ideas emerge. Examples of this include how the acceptable language for certain mental health disorders has changed over time with greater social understanding. Even fairly settled vocabularies such as 'the capital cities of the world' have to be updated when countries rename or move their capitals.

Quite apart from simple factual inaccuracy, there may be wider societal issues of bias and terminology choice to consider in many domains. Some organisations are building vocabularies that allow for more than one preferred label, because there is no one accepted name for say, a mountain or piece of colonised land.

For these kinds of reasons, it is always advisable to have in place a role (or a part of someone's role) to oversee the taxonomy. Yet even this may cause a problem after a taxonomy is deployed. This is because the taxonomist (assuming there isn't a wider team) becomes a Single Point of Failure. When they change companies, retire, or just move job inside the organisation, their knowledge and enthusiasm may not be replaced. (Of course, many companies do keep their taxonomist role filled, but as an external taxonomist I am more likely to see those places where they haven't done this properly!)

Projects can fail even if those in charge fully understand the value of taxonomies in supporting broader business objectives. Managers can have an ambitious vision for using knowledge graphs to power data-driven products and services, yet still not care enough about the quality of the underlying vocabularies and structures.

Models depend on agreed and shared definitions of the types of entities or things the organisation cares about. For example, in one project I was trying to model agreed definitions for digital asset types (e.g. video trailers, background images, call-to-action text). Everyone I talked to agreed there was a need to standardise and reduce the amount of types people were creating. This was because there was rampant inefficiency, duplication of effort, poor communication between production teams and poor findability. Without buy-in from the people overseeing these teams, the processes and systems won't change, and new vocabularies and models won't be adopted. Eventually this may either lead to remediation work being commissioned, or in the worst-case scenario, a total loss of business confidence in using taxonomies at all.

Longer-term sustainable maintenance

Helping others to understand that business-as-usual is as important as project work is a vital task. The taxonomy must be kept relevant and useful for its end-users, and this only has a chance of happening if care and attention are given to ongoing maintenance.

Governance is an important part of this. This can mean anything from a designed framework that encompasses committees and review schedules, to a pragmatic commitment that one person will be responsible for looking after the taxonomy and communicating with others about proposed changes.

It can be easy to fall into the trap of focusing on the 'shiny new thing' that comes out of a project and to underestimate the value of low-key day-to-day deliverables. These include search log analyses, data analytics insights, tagging audits, reporting on taxonomy change requests, and governance processes. All of these things give precious insight into what's working and what's not and may even throw up new and surprising findings. Maybe a whole section of the website is rarely visited, or users are searching en masse for a term variant that no-one thought of. Maybe a taxonomy term is getting disproportionately tagged against content, and it's only because it's the first entry in the autocomplete list for a common word.

Working across disciplines and business siloes for long-term success

There are strategic and practical things that taxonomists can do to help the long-term success of projects. Getting and retaining business buy-in is arguably the most important. There might be one person who is project sponsor, and this is critical, but it is also important to try to build relationships with others across the organisation. After all, taxonomy (and search too!) is not something that fits neatly into one business function. It crosses disciplines such as technology, content strategy, design and user experience, product management and change management. Senior stakeholders in all of these areas should be supported to understand the value of taxonomies.

It's clear that input does not just come from the taxonomist. Other disciplines can and should be involved, e.g. content designers who understand structured content and markup, or developers who understand tagging beyond a simple view of 'stuff added to a piece of content'. I would like to see organisations treat taxonomy, metadata, search and tagging skills as core to their overall set of digital skills.

By way of analogy, it's increasingly common to observe that professionals who aren't digital or content specialists are expected to contribute content about their particular area to digital workplaces. I would like to see organisations encourage staff to interact more with, and understand more deeply the value of, taxonomies. The best colleagues I've worked with over the years had the curiosity and imagination to deeply understand how taxonomies fitted in with their own work and thus benefitted the wider project.

It's great that things are moving in the right direction (after years of 'is taxonomy obsolete?' blog posts!) But it will be even better when people at all levels of an organisation, and across all sorts of digital disciplines, are fully on board with the excellent work that taxonomists are already doing.

This is a revised version of an article that first appeared in Search Insights 2020.

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 were 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.

This is a revised version of an article that first appeared in Search Insights 2021.

Supporting exploratory enterprise search with an AI/ML-based semantic tool

Marianne Lykke, Aalborg University

Introduction

This article presents a semantic tool based on the findings of a study carried out in 2016-18 at an international biotechnology company with 7500 employees. The study was designed to examine searching practices across different work areas and work tasks. The search application was SharePoint 2016. Search log analysis was carried out on 5854 active users over a four-month period. This work was followed by a survey and interviews with eight frequent users and a demonstration of 19 search tasks.

An open access version of the of the original article can be accessed here: https://vbn.aau.dk/en/publications/the-role-of-historical-and-contextual-knowledge-in-enterprise-sea.

How enterprise searchers search

A study of challenges in enterprise search at a Danish biotech company showed how enterprise searchers used an exploratory 'tracing' search strategy in which they used federated search queries to transport them to the part of the information space that contained potentially relevant information. They then continued by using a combination of browsing and searching based on contextual and historical knowledge about previous research projects, products and organisms, lab results, and key people. Searchers used this search strategy to track and interactively put together a set of information that provided the necessary information. Searchers used historical and contextual knowledge to determine the 'tracing' route through the information space.
Another significant result was how people (current and former colleagues), constituted essential sources for obtaining the necessary contextual and historical knowledge. Searchers might not be able to recall relevant product or project names or numbers, but they could recall colleagues or colleagues of colleagues who had been involved in interesting work in the past. They knew that people from an earlier project group or other business areas could guide them to relevant project names, search terms, or to people associated with a project report description, and that searching the person could lead them to the desired information.

The study showed that even very experienced employees with great search competences had difficulties finding information that they were pretty sure was present in the enterprise system. Not only did they have a hard time finding the information they wanted, but they also had a hard time evaluating and assessing whether the retrieved information was relevant and actually about the desired organism, project, or product. Both products and organisms are available in several refined versions, each with specific properties and names, and for the searchers it was essential to be sure that the information concerned the right variant of the organism/product.

Improving the search interface

The results of this study triggered a development project to improve the search interface where the goal was both to support the tracing search strategy with relevant search terms and paths for retrieval, and to give searchers better opportunities to assess the relevance of the retrieved information.

Based on the result, the company chose to add highly contextual metadata to retrieved information, to make it easy for the information searchers to quickly orient themselves in information relevance and possibly find new search terms and paths to the tracing search strategy. By using Scibite, an AI/ML- and ontology-based text analysis search engine, a combination of internal, company-specific, and external, more general biotechnological metadata was added to the enterprise search system's interface in the form of clickable metadata links in the display of retrieved content (see Figure 1).

When the information searcher clicks on the metadata, different types of information are displayed depending on the metadata category. The added metadata leads for instance to predefined Google searches for companies found in the retrieved content, to lab data reports for retrieved lab study results, taxonomic information about organisms from the NCBI National Center for Biotecnology Information database, and project information for internal R&D projects. When a searcher clicks on the added metadata for, for example, an organism, first a list of related organisms is displayed. When the searcher clicks on the name of the organism, relevant information about the organism is displayed that can be used by the searcher to assess the search result and select the next clue and path in the searching. The type of information is different for the eleven metadata types. Some information comes from local, company-specific sources, i.e., the corporate product database, or from external sources, i.e., well-known nomenclatures or taxonomies such as the NCBI. The associated information was defined manually by the corporate information specialists. Both the metadata links and the associated information provide important contextual and historical information. At the same time, it requires knowledge about both context and search to be able to understand and use the assigned metadata knowledge. The assigned metadata was furthermore used to optimise the ranking algorithm, 7% when metadata was found in the body field and 20% when found in the text.

Improving taxonomies and metadata

The development of the semantic tool was carried out in collaboration with Scibite (https://www.scibite.com/), which also provided several of the taxonomies that formed the basis for the identification and assignment of metadata. In practice, however, it turned out that these taxonomies, developed to represent a professional domain in general, were not sufficiently specific in relation to the company's use of and approach to categories. Therefore, internally developed taxonomies were primarily used, and even these had to be developed to achieve a sufficient quality. In practice, this meant that some categories had to be omitted, as they resulted in too high a match and thereby noise in the metadata assignment.

A related problem emerged with the assignment of people metadata. People names were not a problem in themselves, but initials caused problems in the ML processing, as they could be confused with other meanings. The quality of assignment of people metadata was therefore not satisfactory and people metadata unfortunately had to be omitted despite its great importance in search. Another issue was the many variations of a particular organism. As explained earlier, the different variations represented different properties, each with a unique name and in principle easy to distinguish, but since it was the individual researcher who had been responsible for the naming, the naming of organism variants posed a significant problem in the implementation of the semantic tool. As a consequence, the company had to initiate a new development project in order to be able to automate the necessary maintenance and development of the taxonomic semantic content, both in order to manage resources and costs, and to ensure the quality of metadata assignment.

Evaluating the tool and future developments

As shown in Figure 1, the information searcher receives limited information in the display. Another big question is if the user can make an informed judgement about which piece of content and which added metadata with associated information best serves their information requirement, guide them through the search, and ensure the highest possible level of overall search satisfaction.

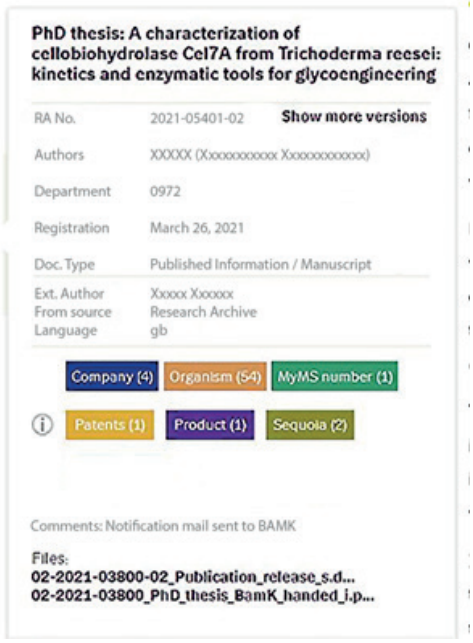


Figure 1: Display of retrieved document with metadata with semantic support.

Therefore, the company has initiated an evaluation of the new feature to examine whether it is understandable to the user, and whether it delivers relevant results and missing clues to help move the user forward in the exploratory, tracing search strategy. A relevant question in the evaluation is also the degree of automation of the information and search support. The developed feature is developed on the assumption that users themselves must control the search process. It is difficult to determine and automate fixed paths, because all searches originate from individual, specific work tasks. The first study identified four main types of search tasks, but in practice, each search task is special, and different combinations of contextual and historical knowledge is required to be able to dive into the enterprise search system and find relevant information.

The evaluation has been initiated and consists of an initial interview study with five frequent users, where knowledge is gathered about the users' search behaviour, use of assigned metadata and their opinion on the new feature. The interview study then forms the basis for a larger questionnaire survey, which is sent to the entire user group to expand and quantitatively confirm the results of the qualitative interviews. The plan is to get insight into general trends and attitudes and to gain further insight into details in user behaviour. The two studies will be followed by participatory workshops, where relevant stakeholders, information seekers, information specialists and IT developers will gather to discuss the findings from the evaluation. The plan is to gather knowledge for the development of AI/ML-based semantic tools to support complex, exploratory searches in enterprise search. Specifically, the intention is to learn about user behaviour and satisfaction to inform the user interface design and further development of the semantic tool.

Communicating enterprise search for success

Agnes Molnar, Search Explained

Enterprise Search is often considered as a "plug-in-and-use" tool: install, deploy, do some configuration - and then it's done, ready to go, and easy to use. Using it does not require any effort, users simply go there, use its capabilities – and are, of course, amazed by it instantly.

However, the reality is different.

Enterprise Search is complex, and can therefore go wrong. And whatever can go wrong – goes wrong. There are many great posts and articles about the complexity of implementation (for example, on CMSWire, Intranet Focus, Search Explained).

The importance of communicating enterprise search is a success factor that is often underestimated. In this article I will discuss the processes and soft skills needed to successfully communicate enterprise search to the organisation.

Why?

The first question to answer is why do we need to communicate Enterprise Search? Isn't it obvious enough? Shouldn't it be easy enough to use?

My experience shows it's not.

Enterprise Search is always a complex system. This complexity is caused by many factors:

- Enterprise search is used by many people across the organisation, in various locations, business units, job roles, etc. Users might also speak multiple languages, have different cultural and educational backgrounds, etc.
  Every user is different – and Search must satisfy them all.

- Enterprise Search is connected to multiple, various systems, with different security models, information architecture, metadata, etc.
  Every system is different – and Search must connect them.

- Enterprise Search has various types of content in the index, with different structures, different formats, and different ways to organise them, etc.
  Every type of content is different – and Search must work with them all.

- Permissions also play an important role in Enterprise Search: search results are security trimmed, and users can find only those documents they have access to.
  Every user has different permissions, and group memberships across the organisation's systems – and Search must respect permissions.

The result is always a multi-dimensional complexity – no surprise that using a complex system of Enterprise Search can be too complex for the users. They need support, education – as well as good communication along the journey.

When?

The next question is when we need to communicate Enterprise Search? Can it be too early? Too late?

Do we need to start involving our users before the rollout? Or even before the detailed planning phase? Or is it enough to let them know when we do the rollout?

Also, what do we need to communicate before, during, and even after the rollout – if anything?

Before

Rolling out a new Search application or a significant update requires preparation. And it's much more than an IT challenge to solve.

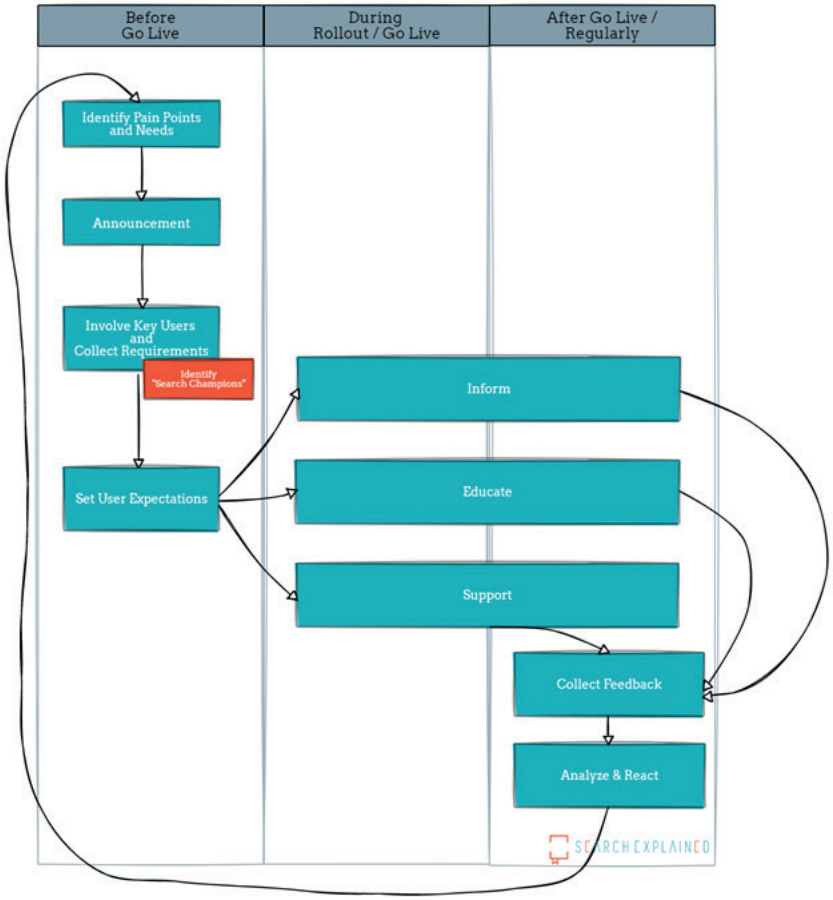


Figure 1: Communicating search

Let's suppose that every search project (whether it is a new implementation or enhancement "only") begins with identifying the pain points and needs. As soon as the decision to go ahead with a search project has been reached, your communication plans also need to be made and ready.

Some organisations make an all-company announcement to inform all employees about the "big news". Others prefer to keep it secret until the implementation is done, or at least enough is done to be demonstrated.

In any case, selected end users should be involved when collecting the detailed requirements. Why? Because they're the ones who know what they really need. They're the ones who know how they would use the relevant results. They're the ones who know what is relevant and what's not.

Of course, management always has a view about search and findability. They always have needs and expectations.

But so do the end users.

Search is always a special application, because it's used organisation-wide, across business units, divisions, departments, and all the job roles. It needs to provide help to everyone. This is why search requirements have to be identified from multiple angles.

Selecting the end users for requirements gathering needs to be done carefully. The Internal Communications Team might have a good idea of who to invite, but these people also need to be asked who they would like to invite. This way, we can make sure that the group of end users involved in requirements gathering is not limited to the ones who are directly known by the Comms team.

While interviewing these key users, doing surveys and workshops, the facilitator also has to watch for exceptionally engaged participants who will be good candidates for becoming Search Champions or Ambassadors later (see below).

Internal Communications has one more very important responsibility: to prepare for the rollout, communicate, and set the end user expectations. The primary role of communication in this phase is to prepare the end users for what's coming and what they can expect. It's important to define and explain the goals of the project (or this phase), the features and/or enhancements to be implemented – as well as the boundaries, and what's not included now.

This way, the end users will have realistic expectations – and if the communication is done well, they'll also be eager and excited by the time of rollout, which can help a lot with general user adoption as well as the overall success of the project.

During the rollout, and after

During the rollout phase, what the users need is three-dimensional:

- Information
- Education
- Support

Information – Again, communication needs to be clear, easy-to-understand, and focused. What the users can expect, when, and how. Where they can get education and support if needed. Who they should contact, and how, if there's any problem or question. (Remember the Search Champions, too – see below.)

Education – Users need search education. Maybe just quick videos or "how to" posts. Maybe formal and informal training courses (online or in-person). Maybe other formats – there are many options. But it has to be clear where and how they can get education about the in-and-out features of search.

Support – Also, users need to feel safe. We have to make sure we communicate clearly where and how they can get their questions answered. Whether it's a central e-mail address, or a channel in Microsoft Teams, or a forum, or any other format – it has to be clear, and accessible for everyone.

(And you need to monitor the questions there, to update the informational and educational materials regularly.)

It's important that all the communications reach everyone in the organisation, regardless of the location, business unit, division, department, or business role of the user. Therefore, we need to create a multi-dimensional communication framework.

How?

Every organisation has its own methods of internal communication. The primary tool may be the Intranet, or email, or Microsoft Teams, or even flyers and posters (in physical offices). There are many more tools and, of course, most organisations use a combination of these.

The most important thing is to know your users and know where and how you can reach them.

Search is considered to be self-explanatory, but it is not! Users need a clear message; they need to understand the (new) features and capabilities. If the message is not clear, or the communication is not visible, search will be under-utilised – and users will never be satisfied with it.

(The message "search sucks" can spread at the speed of light in any organisation. Go to the coffee machine and listen. You'll always find someone complaining about search. Your official communication processes need to compete with this!)

Search Champions and Ambassadors

Besides Internal Communications (and IT), there is one more group of users who can do a lot for the success of search: Search Champions and Ambassadors.

These users work in various job roles across the organisation. And they're engaged. They're excited. They're eager to learn. They're happy to share feedback. And they're happy to be your allies to communicate with the end users.

Keep your eyes open and find potential Champions. Educate them. Socialise with them. Engage them. Support them. And they'll amplify your message. They'll be where the end users are and will support them in-context and on-the-spot. They'll be at the coffee machines and will notice all the feedback – good or bad. They'll be your formal and informal partners to achieve search success.

Conclusions

Enterprise Search can be complex, but it doesn't have to be over-complicated. Users need to have a good understanding of how they can get the maximum out of it – and for this, you need to communicate as soon as possible, as often as possible - and in any way possible to make sure you reach as many users as possible. Involve Internal Communications early – that's the best you can do for the success of any Enterprise Search implementation.

Vector and neural search – a guide to the new frontier

Eric Pugh and Charlie Hull, OpenSource Connections

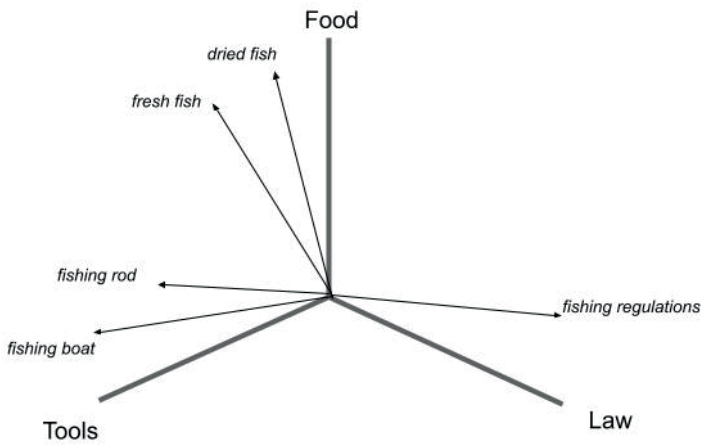
With thanks to Dmitry Kan at Silo.AI for his expert contributions.

If you've been following new developments in search, or noticed a new buzz from search vendor marketing, you will have heard the terms Neural Search and Vector Search. In this article we aim to make it simpler for you to understand this emerging and important field. We'll try to demystify and explain the terms, but importantly also help you decide whether you should be paying closer attention to this trend and how your business might benefit.

What is vector search?

Firstly, what exactly is a vector in the context of search? Let's start by considering how traditional search engines work: essentially they build a data structure much like the index at the back of a book, except they collect and count every single word in the source material. This index makes it very easy to look up relevant documents - if a user types 'fish' as a query, we can quickly find out which documents originally contained that word. To order the documents in terms of relevance, we count how many times 'fish' appears in each document - and indeed across the whole set of documents so we know the local and overall rarity of this word. These calculations are the basis of the Term Frequency/Inverse Document Frequency (TF/IDF) formula used since the 1970s in most text search engines, which works very well in most cases.

The trouble with this approach is that at no point does it capture the meaning or context of the word 'fish' - or indeed the difference or connection between the phrases 'fresh fish' and 'fishing rod', or a recipe for fish cakes and a legal text about regulating the size of fish catches. It's also limited to searching text and can't easily be used for other types of data like images. Luckily, some recent developments in text analysis (like many innovations in search, these were created at large web search companies) have made it much easier to capture the meaning of a piece of text. Machine learning algorithms called Transformers, including BERT and its variants originally developed at Google, can be trained to recognise context and meaning. The result of this training is a mathematical model of a particular language or content area. For the training to work, it needs a large amount of data in this language or context and to indicate success or failure. Feed your source document into a Transformer and the output is a tensor - a multidimensional vector representation for each word in the document, capturing how the model predicts its meaning: note how vectors for similar concepts are close together (this example is of course only three dimensions):



Other machine learning algorithms can be used to extract key features of images, sounds or other types of content as vectors. As these algorithms are often based on neural networks, applications using them are often described as Neural Search engines.

The hard part is figuring out what to do with this huge and complex data structure and indeed how on earth to do it at scale - each tensor can have hundreds of dimensions. Widely used search engine libraries such as Apache Lucene didn't even have a way to store this kind of data in their index until recently, let alone a mechanism for using it for finding which documents match a query. Max Irwin provides a great example of how this matching might be done in a blog from 2019:

> "Commonly, the approach is to use a nearest neighbor algorithm. This takes two or more vectors, and calculates the distance (or similarity) between them in an efficient manner. To use a simple example, let's say we've got two people – Alice, standing 5 meters away from a water fountain, and Bob, standing [in a line] 7 meters away from the water fountain. Who is standing closest to the fountain? Alice of course. Now Carrie joins the group and stands 8 meters from the fountain. So who is Carrie standing closest to? Alice or Bob? Well, that's easy – it's Bob. That's similarity with a vector of one feature. We can also have a vector of two features, such as latitude and longitude coordinates, and use that to calculate who is standing closest to the fountain and who are standing in groups close to each other."

Actual algorithms and approaches now appearing in search engine software include K-Nearest Neighbour (KNN), Approximate Nearest Neighbour (ANN) and Hierarchical Navigable Small World (HNSW) graphs (a form of ANN). There is a large collection of software providing vector search capabilities and now many companies providing this as freely available open source or commercial software - here's a comparison of some current contenders. Recent releases of the Elasticsearch and OpenSearch search servers (both based on the open source Apache Lucene) include vector search features (Apache Solr, also based on Lucene, isn't far behind), whereas the open source Vespa engine natively combines text and vector search at scale.

New approaches to old problems

Let's first consider why this new vector-based approach might be useful to solve some common problems in search.

Similarity search

A common use case is to find items that are similar to each other in some way. Given a source document, traditional text search engines can easily find other documents that contain some of the same words or phrases, but this approach can lack accuracy. Vector search can power a more accurate similarity search based on meaning. Similarity search can also be useful when the intent of the query isn't clear.

Question answering

If someone asking a question of an expert system doesn't use the same terms as the system contains, providing a useful answer can be difficult. Let's consider healthcare: if I am not a doctor, I may not know that a 'myocardial infarction' is a common cause of 'heart attacks' - but the automatically extracted meaning of these two phrases could create vectors that are close to each other. My system can then answer a question like "What is a myocardial infarction?" with plain language for the non-medical professional. Models like ColBERT are actually trained on pairs of questions and passages of text, so they work on a much more refined level than keyword search systems that work on the whole document.

Multilingual search

Building language models enables vector-based matching between a query in one language and a document in another. Historically, building text analysis systems for each new language almost duplicates the effort of the first language, leading to few economies of scale. Moving into new languages can also be difficult as you may lack behavioural data for this language. It is also possible to build a shared multilingual model to avoid any need for direct translation, allowing for a query in one language to search content in another (here's a great demo).

Multimodality

A multimodal search includes different types of content at the same time - for example text and images. As above, vectors can be used to capture features of both textual and not-textual data and then used at the same time to search images directly with a textual query - something that has not been possible before now.

Pushing the limits on search tuning

Tuning an existing search engine is already a constant and complicated process - the complexity, size and richness of the source data is always growing and the number of 'knobs and dials' that can be adjusted also increases. Vector-based approaches allow you to capture this richness and help you establish connections between documents in the collection and between documents and queries in a much more expressive way. It's also worth noting that machine learning approaches such as Learning to Rank (a way to automatically 'learn' the best way to re-rank search results initially provided by a traditional text search engine) are now commonly used in search applications - so ML isn't an entirely new tool for search engineers.

Traditional text search is often used as the basis for a search engine but can be enhanced by this vector-based approach in a two-step process - text search is well understood, scalable and fast, whereas vector search can be hard and difficult to scale. Approaches such as doc2query can even be used in a single stage and beat traditional approaches in some cases according to a recent paper.

Picking a vector search engine

Which route you choose will depend on several factors including:

- Which search platform are you currently using - does it have vector search features or can they be added?
- How much can vector search be a separate add-on to your existing search or do you need a completely new approach?
- Are you able to install and run these complex technologies in-house or would you prefer an externally hosted solution?
- Do you have – or can you acquire either by hiring or training – the new skills in machine learning and data science required?
- Are you prepared to invest in relatively new cutting-edge technology, perhaps from a startup, or would you prefer a more established player?

There is an increasing amount of information on the options available, including the Vector Podcast hosted by Dmitry Kan and talks from OSC's Haystack LIVE! Meetups. As ever, bear in mind that marketing from a vendor may be biased towards their solution or method, and proofs of concept (POCs) using your data and use case are a vital tool. Traditional text search is well known and understood and extremely powerful when correctly applied, and the benefit from vector search approaches may be limited.

The cost of modelling

One thing we rather glossed over above is the cost of training a machine learning (ML) model on your particular content set. Although publicly available pre-trained models are a starting point (for example, you might find one trained on English language news content) these always need to be fine-tuned for your use case to achieve an acceptable level of quality. How effectively this can be done will depend on the data you have available and you will need lots of training data to be successful. A model you find may not easily transfer to your context.

If you haven't before, you are now going to have to consider the cost of running machine learning models in your organisation, including the time to create and debug models, the cost of running computationally expensive training cycles and how to rapidly and effectively deploy new models at scale (the operations process commonly called 'MLOps'). You should be constantly training all of your models on a regular ongoing basis and should think of ML as a manufacturing process that is continually running, not like software engineering which is typically project based.

Dmitry Kan has some ideas on how to measure the results of this investment:

- Each model impact should be evaluated on its influence on the business KPIs e.g. transactions per user, gross merchandise value for e-commerce, grounded in the cost for running ML in the company
- Model level metrics should roll-up to a system KPI. Example metrics are accuracy, Precision, Recall, F1 or more specific like Word Error Rate (WER)
- Since MLOps is an organisational problem, measure the percentage of time spent on infrastructure debugging per ML researcher, average time from training to production deployment, model query monitoring and knowing what to monitor is the key to success - such as drift caused by data change for any reason, including bugs in data

Note that some technology providers promise to handle much of this for you - of course, you will still need to supply them with content and data to work with, and you will need confidence in their ability to deliver, their overall approach to machine learning and indeed their models.

Parting thoughts

Neural and vector search promise to deliver huge leaps in search quality and can enable use cases that have traditionally been difficult to deliver with traditional text-based approaches. However, there is a plethora of approaches, technologies and companies and it can be hard to choose between them. Effective vector search is hugely reliant on high quality training data and may also require significant investment in machine learning and operations, unless you are able to trust third party providers to do the heavy lifting for you. The future of search undoubtedly contains vector and neural approaches, but remember that all that glitters is not gold!

Links

- The Haystack LIVE! Meetup features many recorded talks on vector search https://www.youtube.com/playlist?list=PLCoJWKqBHERvPMypkyJPoPtZKvcPJ9KiD
- Dmitry Kan's blog https://dmitry-kan.medium.com/
- Vector Podcast https://dmitry-kan.medium.com/vector-podcast-e27d83ecd0be
- Vector Databases and Neural Search, a presentation by Dmitry Kan https://www.slideshare.net/dmitrykan/vector-databases-and-neural-search-251033449

- OpenSource Connections blogs on BERT
  https://opensourceconnections.com/blog/2019/11/05/understanding-bert-and-search-relevance/
  https://opensourceconnections.com/blog/2019/12/18/bert-and-search-relevance-part2-dense-vs-sparse/
- OpenSearch / Elasticsearch KNN/ANN announcements, blogs
- Neural Search in Solr https://sease.io/2022/01/apache-solr-neural-search.html
- Announcing ANN in Elasticsearch https://www.elastic.co/blog/introducing-approximate-nearest-neighbor-search-in-elasticsearch-8-0
- kNN in OpenSearch https://docs.aws.amazon.com/opensearch-service/latest/developerguide/knn.html
- The Haystack Search Radar is an attempt to track innovations in search https://haystackconf.com/radar/
- The free Relevance Slack (over 2300 members to date) has a #vectors-in-search channel http://www.opensourceconnections.com/slack
- Database technology evolves to combine machine learning and data storage https://venturebeat.com/2022/03/07/database-technology-evolves-to-combine-machine-learning-and-data-storage/
- From doc2query to docTTTTTquery, an MS MARCO passage retrieval task [1] publication, Rodrigo Nogueira1 and Jimmy Lin2 https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery-v2.pdf

Searching fast and slow

Tony Russell-Rose, 2D 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?

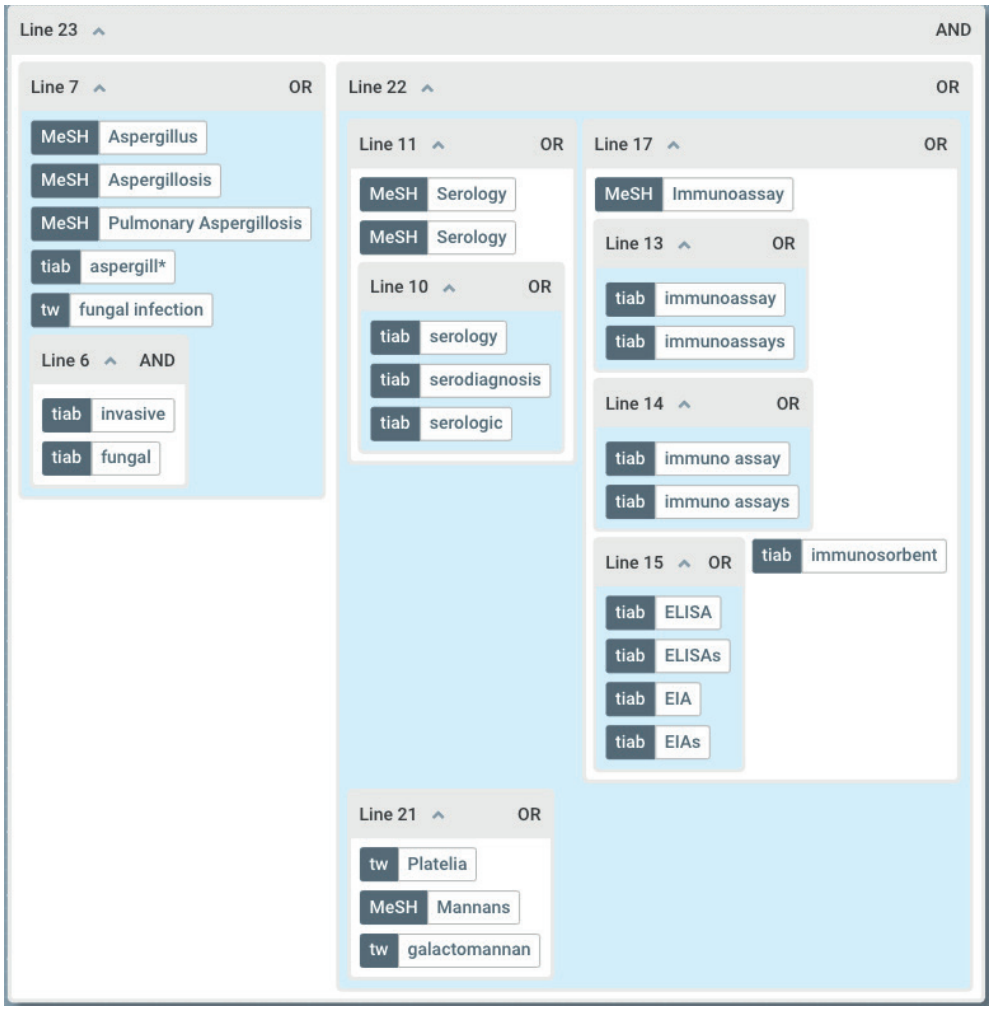
Clearly all these questions are answerable, albeit more so to the trained eye. But the point is that the answers are not directly visible. Instead, we must proceed through a sequence of steps: we must first retrieve from memory a method for interpreting line-by-line searches, and then implement it. In 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:

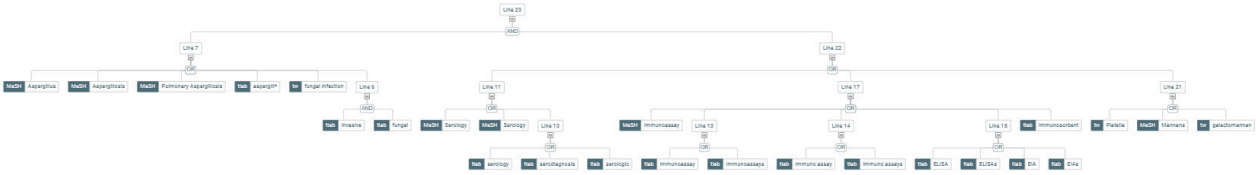


This visualisation reveals that our search strategy from earlier consists of a conjunction of two disjunctions (lines 7 and 22), the first of which articulates variations on the fungal infection concept, while the latter contains various nested disjunctions to capture the diagnostic test (serology) and 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.

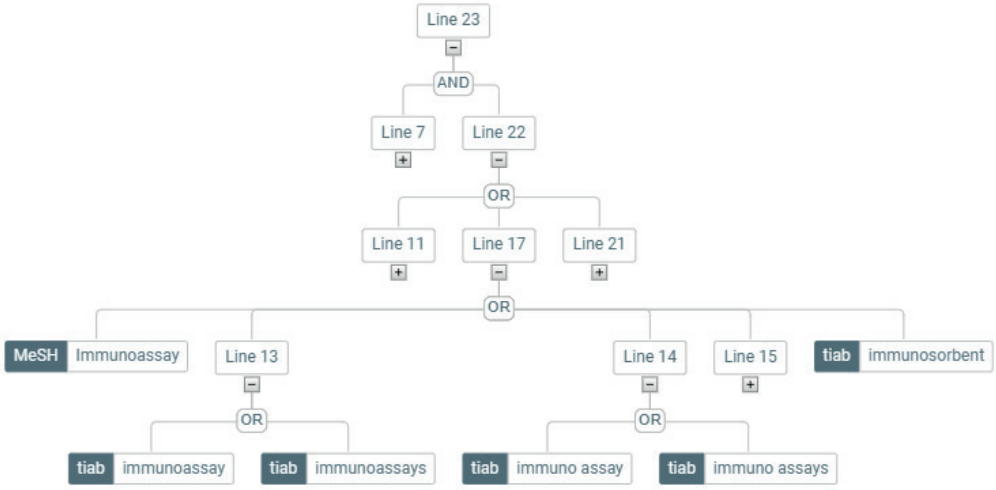
Tree view

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:

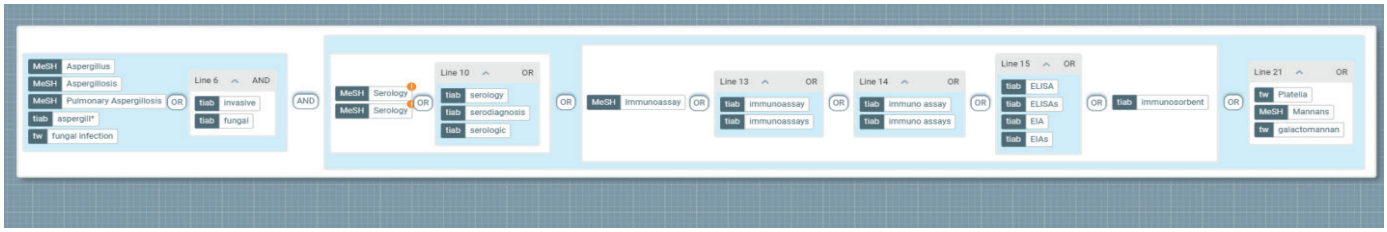


Like the Nested view, the Tree view maps conceptual hierarchy onto physical hierarchy, but in a manner that emphasises branching over containment. But is hierarchy the most important aspect of a search? With that in mind let's examine a third view.

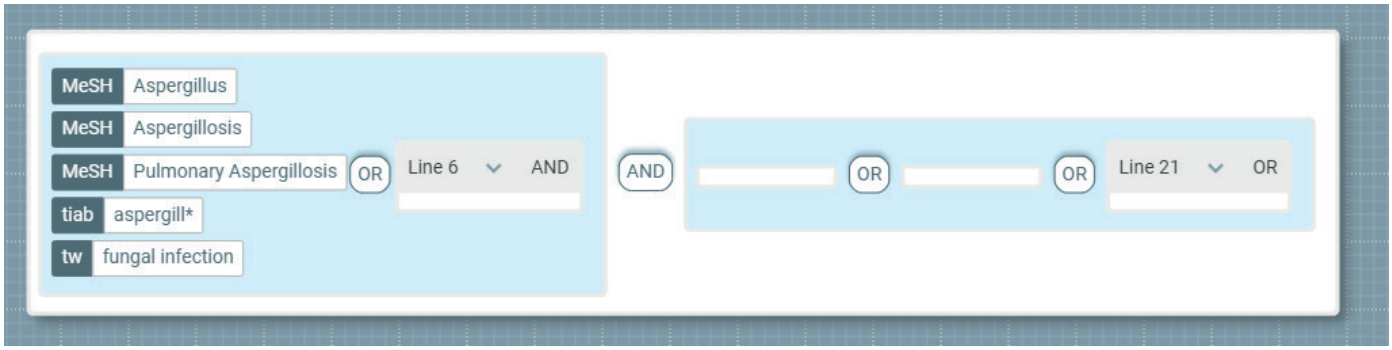
Inline view

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.

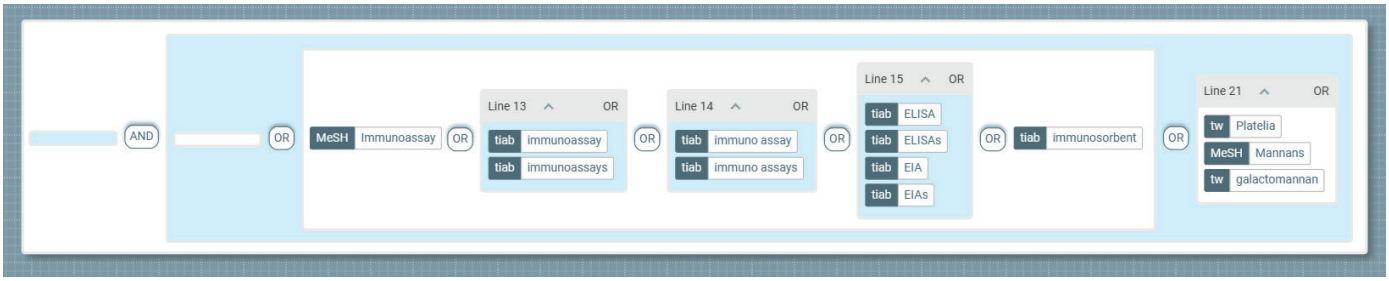
Like the Nested view, the Inline view maps conceptual hierarchy onto physical containment, but this time in a manner that aligns groups along a common midline, giving rise to a natural left-to-right reading:



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 article 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.

This is a revised version of an article that first appeared in Search Insights 2021

Learning about machine learning

Martin White, Intranet Focus

1. Implementing machine learning in the enterprise

The business and technology case for using machine learning to optimise web search and e-commerce search applications is very strong. Both generate very large quantities of search log data that can be used by AI/ML applications to optimise the results presented to a user.

The objective of this article is to stress the importance of due diligence in matching the claims from search vendors against the current and potential requirements of the organisation and its employees.

Enterprise search vendors have been assiduous in presenting the benefits of machine learning in achieving the highest possible levels of search satisfaction but provide very little information about the processes involved in implementing their machine learning solutions. The consulting firm Deep Analysis, which specialises in AI, has suggested that in many cases what is taking place is 'AI washing', in which existing technology is given the appearance of having been transformed into an AI application.

This article highlights the challenges that vendors and customers face in implementing AI/ML models for enterprise search use which is different in so many ways to web and e-commerce search.

2. The long tail

In 2021 Professor Marianne Lykke and her colleagues published the first ever detailed analysis of the patterns of enterprise search use inside a single organisation (see page 40).

The analysis of the distribution of searches by category is shown in the table below

People search	63.4
Quality	4.4
IT	2.6
HR	2.2
Product	1.9
Finance	1.1
Facility services	1.0
Sourcing	1.3
Intranet	0.8
No top category	17.2

The number of queries for people is usually invisible to a search analytics team because the number of queries about an individual employee is so low that it comes at the end of the long tail of queries ranked by frequency of search.

Another example of the extent of the long tail is shown in the table below, which shows the ranked number of queries per month in a global company with just over 100,000 employees.

At Rank	Number of queries/month
10	36133
20	19282
30	14453
40	11865
50	9455
100	6218
500	2307
1000	1361

The query term at position 1000 was 'Sentinex'. Although the number of queries for this term are only just over 1000, the context is important. Sentinex was a new product that was about to be released. At this stage a small group of engineers and sales teams were making business-critical decisions about the product and needed to be able to check a range of product engineering data in formulating a release strategy, much of which was associated with SU45/17/MT, which was the project code designation. The result set for Sentinex contained a mix of Word, PDF, Excel, PPT, CAD, png and video files, and the search support team had to put in a significant amount of work to link data associated with SU45/17/MT with Sentinex.

This example is provided to show the complexities of enterprise search that cannot be resolved with the limited amount of data on either SU45/17/MT or Sentinex.

3. Defining objectives

There are four fundamental questions to be answered at the outset of any discussion about the potential benefits of implementing a machine learning model.

- What is the problem that needs to be solved?
- Is ML in principle a potential solution?
- How will the increase in search performance be assessed?
- Will the effort be justified by the performance increase?

Until the answers to these questions have been agreed and documented there is no benefit in beginning a potentially open-ended discussion with a search software vendor.

4. Training ML applications

In the initial meeting with a prospective vendor you should be able to review the documentation of the development process, and in particular understand the content of the data sets on which the vendor trained its ML model.

Implementation of a machine learning application involves using a Training set, a Validation set and a Testing set. A blog post from Label Your Data provides a good presentation of data set management.

In principle these data sets should be a close match to the content of your organisation but the problem all vendors face is where to source the training data for their development team. There are many data sets available but none even remotely representative of enterprise collections. For some years there was a TREC Enterprise Search track that used public data sets that were considered (without any specific justification) to be representative of actual enterprise search collections.

At the same time it is important to understand who developed the ML algorithms and how they have been tested. The 'who' is important as it is possible to use a contractor to undertake the testing work, and it could be that the rights to the algorithm are not owned by the vendor itself.

Just as important is understanding how the outputs of the training and validation stages were assessed. It is quite common to take a data set and run it through a BM25-based relevance engine to come up with an initial test collection, but that inevitably raises the issue of who made the relevance judgements and on what basis. This is where the issues of bias and fairness come into play. Should the relevance judgements be made by a team at corporate HQ or in subsidiaries that are generating, searching for and using the content?

5. Sizing the data sets

The unknown in machine learning development is how large the data set needs to be to achieve an acceptable level of performance. In the case of enterprise search it is not widely appreciated that in terms of query volume 'people search' is probably an order of magnitude greater than any other category, and might represent at least 60% of all queries. The remaining queries are being generated from perhaps 80% of employees covering a very wide range of topics, some of which require high precision in the results list and others where high recall is essential.

Then there is the need to factor in the diversity of largely un-curated content in an enterprise environment. What is certain is that brute-force unsupervised ML models are totally inappropriate for enterprise search requirements.

6. The role of the vendor

In theory a vendor should have some prior experience to share of quantifying data set size but in practice the vendor may be unwilling to do so because the obvious next question is to ask to talk to other customers of the vendor to gain their feedback on the process.

Protective marking schemes also need to be taken into account. Should the test collections consist only of content that is accessible to every employee, or should there be a number of collections at various levels of access? It could be that content that is not widely accessible is of especial importance to a comparatively small group of managers.

It is not just the number of documents in the test data sets that is an important factor but also

- how the task of learning was shared between the vendor, the customer and potentially an integrator
- how long the process took
- was an A/B test carried out to assess the improvement in retrieval performance in terms of either precision or recall.

Enterprise search solutions inevitably develop, expand and contract over time as the business interests of the company change, new subsidiaries and business initiatives are created, and other businesses are acquired. All these will change the requirements and expectations of the search user. This raises the issue of the support required from the vendor and/or integrator once the initial implementation has been delivered.

7. Assessing relevance

Another factor to consider is the potential bias of search results based on the choice of data sets, the choice of queries, and the decisions on what is, or is not, relevant. Judging relevance on the basis of a snippet is not a reliable approach as the snippet itself may not be representative of the content item. Algorithmic bias is now recognised as being a major issue to address in managing AI applications.

To overcome this problem, it will be necessary for the search team and the relevance judges to look through the entire document, which will be a time-consuming process that will still raise the issue of whether the relevance judgement is free from bias.

The bias is especially problematic when ML is used to deliver personalised search results and/or recommendations. The search user may not be aware of the basis on which the personalisation was carried out in terms of both the factors and the weighting of the factors in the ML basket.

8. Explainable AI

Over the last couple of years the need to develop, implement and govern explainable AI policies has become very evident. A review paper published in late 2021 that provides a good overview of the issues and potential solutions comes with a bibliography of almost 200 research papers and other documents. The initial question with explainable AI is who is the explanation written for? Different stakeholders need different explanations of a technology which tends to get submerged in jargon and abbreviations. IBM has published a very helpful 'question bank' to support a discussion about explainable AI. There is a video summary of the approach IBM has taken.

9. Summary and conclusions

There is no doubt that AI/ML models offer some significant benefits in enhancing the performance of web and e-commerce search. With enterprise search the benefits are much less obvious even though many search software vendors indicate that AI/ML methods can transform the enterprise search experience. Very little, if any, evidence is presented to justify this claim.

That is not to say that there could be benefits, perhaps for specific groups of search users across some equally specific content types. Defining these groups and content types is going to require detailed consideration by both the search team and line-of-business managers.

Managing multilingual and cross-lingual search

Martin White, Intranet Focus Ltd

Inside an organisation it is easy to overlook the importance of the range of working languages in use even in a single country, let alone in other countries. Although English is often assumed to be the universal language of business this is a perspective convenient to people with English as their primary (and probably only) language.

Although there has been a significant amount of research and development into Natural Language Processing (NLP) this has tended to be at the language structure level. It is also important to understand the overlaps between text mining, text analytics and natural language processing. However, it is important to acknowledge that speakers and the content they produce and require vary not just by language, but also by culture. Although language and culture are tightly linked, there are important differences. Analogous to cross-lingual and multilingual NLP, cross-cultural and multicultural NLP considers these differences in order to better serve users of NLP systems.

Global multi-national companies inevitably operate in multiple languages (Thomas 2007):

These can be categorised into

- the parent company language
- a common corporate language
- multiple local (foreign) languages.

The parent company language (PCL) is often the language spoken at the headquarters company and is often an official language of the nation in which the parent company operates. The common corporate language (CCL) is the default global operating language. In the case of US and UK companies the PCL and the CCL are English. In the case of a German-owned company the PCL is German and the CCL is often English.

However, corporate language is much more complex. Table 1 shows the content language distribution in a global company which has its headquarters in Germany.

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

Table 1: Language diversity in a multinational pharmaceutical company

The high percentage of content in English is not unexpected but it is important to note that only 24% of employees have German (based on data of their regional location) as their primary language. This means that 76% of employees are potentially searching

for information in their second language, possibly with a restricted command of the language when it comes to choosing synonyms to improve the results from a query or fully understanding a search result snippet.

The percentage of content in Spanish and Portuguese shown in the table does not reflect that there are differences between the European (Castilian) and South/Latin American versions of these two languages. As one example, a mobile phone is teléfono móvil in Spanish and cellular in Latin American Spanish and computer is likewise ordenador or computadora. The differences are more marked in the spoken language and in social media.

Looking at the table the inclusion of Polish, albeit as just 1%, is a result of the company having significant administration, finance and IT operations based in Poland and the internal language of these departments is Polish.

Companies are now starting to develop formal language policies which take account of the levels of language competence and the impact of technology in supporting a range of working languages on a global basis. Implementing these policies requires there to be training and support in working in two or more languages and monitoring the success of these initiatives on achieving corporate objectives.

Multilingual and cross-lingual search

It is important to distinguish between multi-lingual search and cross-language search, though the term 'multi-lingual' is often used for cross-language search. In a multi-lingual implementation the content is indexed using management tools that are specific to a language, together with appropriate stop words. Most enterprise search applications are able to identify a language at index time and apply the appropriate linguistic tools in the process of tokenisation.

The query itself is parsed and matched across the index to detect the language of the query and then the query is run against the correct index. In effect the search application is language-independent.

However, it is not unusual for documents to be in more than one language, and for these the crawl needs to highlight them on an individual basis for the search team to consider how best to manage them. Job vacancies are a good example, where the formal title might be in the local language, but the position is open across the organisation to employees who have the appropriate language and technical skills. It also has to be appreciated that countries may have more than one official language. Examples include Belgium, Spain, Switzerland and the UK (English and Welsh).

Language detection and index management

The initial requirement is for the search application to detect the language of a piece of content at the indexing stage and then create a valid index with the appropriate stemming and lemmatisation routines. Even this can present problems:

- The document is in a single language
- The document is predominately in a single language but may contain snippets (perhaps quotes from customers) which are in one or more different languages
- The document is in multiple languages but within a specific field (e.g. Executive Summary) there is only a single language
- The document has multiple languages in one or more fields

The issue of variations in fields is often a result of metadata being applied in one language, the text being in a second language (e.g. Chinese) and in addition there is a summary of the document in English.

The goal, although not always achievable, should be to keep languages separate. Mixing languages in the same inverted index can be problematic.

A crucial step in the indexing process is to apply the correct stemming rules for each language. The stemming rules for German are different from those for English, French, Swedish, and so on. Applying the same stemming rules to different languages will result in some words being stemmed correctly, some incorrectly, and some not being stemmed at all. It may even result in words from different languages with different meanings being stemmed to the same root word, producing confusing search results for the user. Applying multiple stemmers in turn to the same text is not advisable, as the next stemmer may try to stem an already stemmed word, compounding the problem.

Mixing languages also has implications for the weighting of terms in retrieval. A short snippet of German appearing in predominantly English text would give more weight to the German words, given that they are relatively uncommon. But mix those with documents that are predominantly German, and the short German snippets now have much less weight.

There are two additional issues to take account of. The first is in supporting searching for the names of employees as these also show linguistic variation. As an example, Portuguese surnames are often presented as maternal – paternal family names, but the reverse is the case in Spanish, and in spoken language the maternal name is often, but not always, omitted.

The second issue is that of entity extraction, especially where locations can have both an accepted global name (Cologne) and a national name (Köln) which in this example has a symbol that requires an extended ASCII code. Documents relating to Germany could be in English but through common use contain various entities that are in a local language.

Query management

User searches also come in three main varieties:

- Users search for words in their main language.
- Users search for words in a different language but expect results in their main language.
- Users search for words in a different language, and expect results in that language (for example, a bilingual person).

Identifying the language of the user's search request is not a straightforward task. Common language detectors are designed for text that is at least 200 characters in length. Shorter amounts of text, such as search keywords, produce much less accurate results. In these cases, it may be preferable to take simple heuristics into account such as the country of origin, the user's selected language, and the HTTP accept-language headers.

Some recent research showed that multilingual search users often tend to construct queries in their primary language. However participants often preferred search result lists in their secondary language when presented with this choice. The reason for this could be that a search user will post a query in the language they are most familiar with

but will have a good enough command of their second language to feel comfortable in understanding results in this second language, especially if for some reason it is likely that these results could be more relevant.

Similarly, when given the choice between prewritten queries in different languages, users often select queries in their secondary language to search for information. These observations confirm that the current "unilingual" view of localised services does not reflect or support multilingual user behaviours and preferences.

User interface options

There are parallels between multilingual search and federated search with regard to the presentation of results from different applications and repositories in the user interfaces. The options have been investigated in depth in the context of a multilingual news search service and these are broadly of relevance in enterprise search.

Cross language search

A much more challenging requirement is for a search to be conducted across multiple languages where the search term is in one language and there is a requirement to match the term in all the other languages. This is cross-language search and requires a multi-lingual thesaurus. A good example is the Eurovoc thesaurus developed for the 23 languages of the European Union and there is a 30,000 word multilingual thesaurus developed for nuclear research by INIS.

To ingest and manage multilingual thesaurus search, software applications require them to be developed to ISO 25964 – the international standard for thesauri and interoperability with other vocabularies. This is a very rigorous standard and compiling a thesaurus that meets this standard is an exceptionally challenging process, especially in science and technology where new terms appear on a frequent basis. As a result, cross-language search applications tend to be used only in large-scale public sector projects or for specialised collections.

In conclusion

Because of the entanglement of language and culture, and often a lack of transparency and understanding of the importance of supporting local languages, developing a strategy for multi-lingual search on an ad hoc basis as requirements emerge is not a sensible approach. There needs to be both a top-down process in the context of corporate language policies and a user-upwards process to identify the operational requirements of users with different language skills.

It is also of considerable importance that the search support team looks carefully at the query terms being used in any of the supported languages to ensure that problems being experienced by employees are recognised. This requires the search team to have the support of employees with an appropriate range of languages.

Search resources – books and blogs

The books listed below represent a core library which should be on the bookshelf of any manager with enterprise search responsibilities. 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 cornerstone 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)
This book outlines the technology of enterprise search, the operational management of search applications and sets out the processes involved in justifying, specifying, selecting and implementing an enterprise search application.

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.

Accenture Search and Content Analytics blog
Beyond Search Stephen Arnold
Complex Discovery Rob Robinson
Coveo Insights Corporate Blog
Daniel Tunkelang
https://www.enterprisesearchblog.com/
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 Explained Agnes Molnar
Sease Corporate blog
SearchStax 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 is a regular column on search written by Agnes Molnar and published by CMSWire.

List of enterprise search software applications

The Europe column indicates whether the company is based in Europe or has an office in Europe, based on the information on the company's web site.

Company	HQ	Europe	URL
Algolia	USA	SaaS	https://www.algolia.com
Amazon	USA	SaaS	https://aws.amazon.com/kendra
Ankiro	Denmark	Yes	https://ankiro.dk/ankiro-enterprise-suite
Appbase			https://www.appbase.io/solutions/saas-search
Aras	USA	Yes	https://www.aras.com/en/capabilities/enterprise-search
Ayfie	Norway	Yes	https://www.ayfie.com
BAInsight	USA	Yes	https://www.bainsight.com
Bloomreach	USA	Yes	https://www.bloomreach.com/en
Bonsai	USA	No	https://bonsai.io/
Capacity	USA	No	https://capacity.com/capacity-enterprise-search/
Cludo	Denmark	Yes	www.cludo.com
Cognite	Norway	Yes	https://www.cognite.com/en/company
Copernic	Canada	No	https://copernic.com/en/
Coveo	Canada	Yes	http://www.coveo.com
Curiosity	Germany	Yes	https://curiosity.ai/
Dashworks	USA	No	https://www.dashworks.ai/
Datafari	France	Yes	https://www.datafari.com/en/
DieselPoint	USA	No	http://dieselpoint.com
Dokoni Find	USA	Yes	https://www.konicaminolta.co.uk/en-gb/software/dokoni-find
dTSearch	USA	Yes	http://www.dtsearch.com/
D.velop	Germany	Yes	https://www.d-velop.com/software/enterprise-search
Elastic	Netherlands	Yes	https://www.elastic.co/products/elasticsearch
Exalead	France	Yes	https://www.3ds.com/products-services/exalead/products/
Expert.ai	Italy	Yes	https://www.expert.ai/
Fess	USA	No	https://fess.codelibs.org/
Findwise	Sweden	Yes	http://www.findwise.com
Funnelback	Australia	Yes	https://www.squiz.net/products/funnelback\
Google Cloud Search	USA	Yes	https://workspace.google.com/products/cloud-search/
Greentree	UK	Yes	https://greentree.co.uk/software/information-access-and-reporting/enterprise-search
Hulbee	Switzerland	Yes	https://hesbox.com/en/overview/at-glance
Hitachi	USA	Yes	https://www.hitachivantara.com/en-us/products/storage/object-storage/content-intelligence.html
Hyland	USA	Yes	https://www.hyland.com/en/platform/product-suite/enterprise-search
IBM Watson	USA	Yes	https://www.ibm.com/watson/products-services
Ilves	Finland	Yes	https://ilveshaku.fi/en/

Company	HQ	Europe	URL
iManage	UK	Yes	https://imanage.com/product-overview/platform/
Inbenta	Spain	Yes	https://www.inbenta.com/
Indica	Netherlands	Yes	https://indica.nl/en/enterprise-search
Infongen	USA	Yes	https://www.infongen.com/solutions/enterprise-search
Intergator	Germany	Yes	https://www.intergator.de/en/solutions-applications/enterprise-search/
IntraFind	Germany	Yes	https://www.intrafind.de/index_en
Klera	USA	No	https://klera.io/
Knowlia	Belgium	Yes	https://www.knowliah.com/
Lucidworks	USA	Yes	http://www.lucidworks.com
Lucy	USA	No	https://www.lucy.ai/about-us
MarkLogic	USA	Yes	https://www.marklogic.com/solutions/enterprise-search/
Microfocus	UK	Yes	https://www.microfocus.com/en-us/products/information-data-analytics-idol/overview
Microsoft Azure	USA	Yes	https://azure.microsoft.com/en-us/services/search/
Microsoft SharePoint	USA	Yes	https://docs.microsoft.com/en-us/sharepoint/dev/general-development/search-in-sharepoint
Mindbreeze	Austria	Yes	https://inspire.mindbreeze.com/
Nalytics	UK	Yes	https://www.nalytics.com/
Netowl	USA	No	https://www.netowl.com/enterprise-search
Nuclia	Spain	Yes	https://nuclia.com/
Onna	USA	No	https://onna.com/enterprise-search/
OpenText	Canada	Yes	https://www.opentext.com/what-we-do/products/discovery
Precognox	Hungary	Yes	https://www.precognox.com/intelligent-search/
Raffle	Denmark	Yes	https://www.raffle.ai/
ResoluteAI	USA	No	https://www.resolute.ai/about
Searchblox	USA	No	https://www.searchblox.com/
Searchunify	USA	No	https://www.searchunify.com/
Sinequa	France	Yes	http://www.sinequa.com
Sirius	UK	Yes	https://www.siriusopensource.com/en-uk/solutions/search
Squirro	Switzerland	Yes	https://squirro.com/
Swiftype	USA	Yes	https://swiftype.com/
TeraText	Australia	No	https://www.teratext.com/about/
Theum	Germany	Yes	https://www.theum.com/cognitive-search-and-knowledge-discovery/
Thunderstone	USA	No	https://www.thunderstone.com/lp/enterprise-search/
Unleash	N/A		https://www.unleash.so/about
Voyager	USA	No	http://www.voyagersearch.com
Weaviate	Netherlands	Yes	https://weaviate.io/
Yext	USA	Yes	https://www.yext.com/
Zakta	USA	No	http://zakta.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 a specific repository, a set of documents, or a section of a document.

Advanced search
The provision of a search user interface which prompts the user to enter additional terms to assist in 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.

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

COTS
Commercial off-the-shelf software.

Crawler
A program used to index documents.

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

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.
(See also Information source.)

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

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

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

Exploratory search
In exploratory search the search goal is imprecise and open-ended and there is no unique single answer that meets the user's information needs and no clear criterion on when to end 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 knowledge graph is 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)
Learning to Rank is a class of techniques that apply supervised machine learning to solve ranking problems by presenting a relative re-ordering 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 groups of professionals (for example, lawyers and patent agents) who spend a significant proportion of their time using search applications, often in situations where high levels of recall are required.

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.

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

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

Truncation
Removal of a prefix or suffix.

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

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

Weight
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
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.).
(See also 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.

Note:
A very extensive glossary on AI and Machine Learning can be found here.