Behind the Curtain: Understanding the Search and Discovery Technology Stack

Patrick Lambe

> The Lion thought it might be as well to frighten the Wizard, so he gave a large, loud roar, which was so fierce and dreadful that Toto jumped away from him in alarm and tipped over the screen that stood in a corner. As it fell with a crash they looked that way, and the next moment all of them were filled with wonder. For they saw, standing in just the spot the screen had hidden, a little old man, with a bald head and a wrinkled face, who seemed to be as much surprised as they were. The Tin Woodman, raising his axe, rushed toward the little man and cried out, "Who are you?"
>
> "I am Oz, the Great and Terrible," said the little man, in a trembling voice. "But don't strike me--please don't--and I'll do anything you want me to."
>
> *L. Frank Baum. The Wonderful Wizard of Oz. Project Gutenberg (2008)*

The Wizard of Oz captures the life trajectory of many technologies, and search is no exception: the technology starts in the public domain, it is appropriated and made mysterious (hidden behind the curtain) by commercial interests, then touted and overblown as magical, eventually the curtain falls, and we can see what has been there all along, and work with it more intelligently as it is, knowing what it's good for and knowing what it's not good for. This is true of search, as it is true of auto-classification. Modern search technology has its roots in Vannevar Busch's seminal 1945 paper "As we may think" and its foundations were developed by research groups at MIT, Cornell and Harvard into the 1970s.

One of the consequences of hiding a technology behind the commercial curtain is that different technologies compete with each other for the same market, and therefore downplay their perceived competitors, or attempt to duplicate the functionalities of their perceived competitors. Search technology has competed with taxonomy work for years – "you don't need taxonomies," went the mantra of one infamous search vendor, "our software will understand your content for you." The curtain fell with the unveiling of Google Knowledge Graph, when it became obvious just how much Google was investing in knowledge organisation structures (controlled vocabularies, taxonomies, ontologies, knowledge graphs) to make its search smarter. The superior performance of open source search engines (and their striking adoption rates) over commercial search engines in solving particular types of problem is, like Toto's leap, another manifestation of the dropped curtain.

Machine classification of content has been another example. It is frequently touted by its commercial vendors as a single technology performing a single "we'll tag your

content for you" function where it is actually a bundle of quite distinct technologies, and rooted in information retrieval research and search technologies developed in the 1970s and 1980s. Machine classification vendors have found themselves competing with search ("we can make your search smarter") and taxonomies ("we can derive taxonomies automatically for you"). The curtain is now starting to fall with large organisations learning how to use the open source toolkits such as the University of Sheffield's GATE toolset, to solve their search and discovery problems.

The "go it alone" propaganda of the technology vendors is damaging because it locks buyers into a single toolset designed for common denominators, with prohibitive customisation and maintenance costs. The "magic black box" propaganda blinds organisations to the need to build capabilities in a range of tools. And the "we do everything" propaganda hides the synergies and capabilities that can be attained by intelligently using search, machine classification and knowledge organisation technologies as a toolkit, with different toolsets to be used in different combinations to solve different kinds of search and discovery problem.

The purpose of this article is to explain the uses and interactions of three technologies that together support search and discovery, so that organisations can learn which tools to adopt in which combination to solve different problems, and they can see what kind of internal capabilities they need to build and maintain to continue solving those problems:

- Enterprise search
- Taxonomy and ontology management
- Text analytics to perform machine classification.

Enterprise search is a technology for delivering useful and relevant results to users exploring a large content base. As a technology it is relatively simple and not especially smart, and to deliver superior results in complex environments it needs to be supported by other technologies. "Complexity" can refer to the diversity of user communities, and to the diversity of the content itself. Search engines by themselves don't understand anything. They crawl and index, process queries and serve up user analytics. To become smart, they need human designed curation tools (taxonomy and ontology tools) and tools for scaling the human tagging of content (text analytics tools).

Taxonomy and ontology management technologies support designing, delivering and maintaining knowledge structures and controlled vocabularies to enhance the search and discovery experience. These structures are based on an analysis of business and user needs together with an analysis of the target content for search and discovery, and they help to disambiguate concepts, capture variations in language and map them as synonyms, and identify relationships between concepts so that searches can be expanded or narrowed. The constraints of taxonomy and ontology management most often relate (a) to staying up to date with new and emerging needs of diverse user sets (which is where search analytics can help), and (b) to scaling the way that concepts can be identified in content and metadata enriched (which is where text analytics and machine classification can help).

Text analytics refers to a cluster of technologies that extract meaning from text and turn it into metadata. Some of the root technologies (crawling and indexing) are shared with search. These technologies are a response to the problems of human inconsistency (people are not consistent in how they apply tags to content) and of the scale of content to be tagged (when it would not be feasible to apply human-curated tagging to large amounts of content in a short period of time). Different technologies are used to supply different types of tagging operation, ranging from identification of known entities such as people, organisations and places, quantities such as money amounts, to the identification of abstract concepts that characterize what a document is "about" in ways that make sense to target users. The constraints of text analytics are (a) that it has a hard time figuring out without intensive human curation or very well structured training sets which concepts are most salient to the most people (which is where taxonomy and ontology management helps) and (b) like search, it requires constant tuning based on emerging user needs (which is where search analytics on user behaviours can help).

These technologies can be extremely powerful when used in combination in support of the search and discovery experience. However, it is important to understand what they are capable of, how they work, and the limitations of what they can do.

Figure A1 shows a detailed view of the three pillars of the search and discovery technology stack, how they work, and their key interactions. We begin at the top of the diagram. Since these technologies are supposed to be deployed in the service of user needs, decisions about their configuration and design need to be made in the context of known user needs, as well as a thorough understanding of the content to be covered. Note that this is a simplified conceptual framework, and it does not fully represent the complexities of the relationships between the pillars, nor all the overlaps between them.

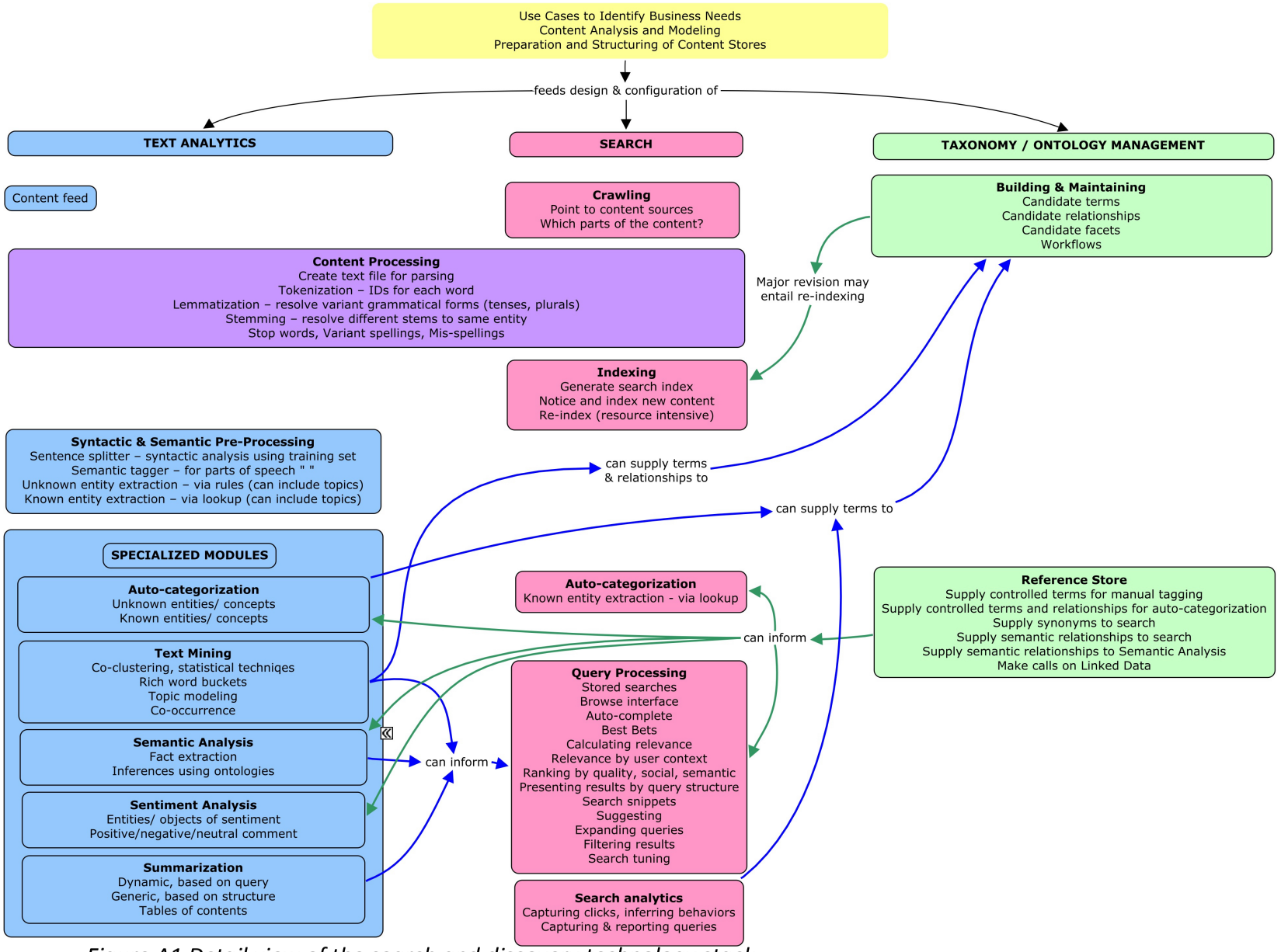

Figure A1 Detail view of the search and discovery technology stack

“Content” can be structured, semi-structured or unstructured. It can be in the form of data, documents, web content, audio visual files, discussions, or questions. “People” can also be considered a form of content since knowledgeable people may also be a useful result in s search on a particular topic. In a content management system a person profile can be used as a proxy for that person, and can be tagged with relevant subject tags so as to surface in search alongside other forms of content.

All three technologies in the stack depend upon a thorough survey of the target content to be covered, and preparation and structuring of the content stores so that the search, taxonomy management and text analytics technologies can access them and operate upon them.

Enterprise Search

Let’s begin with the central pillar, the enterprise search stack. There are six main components in the search stack.

Crawling

The first component is the crawling module. This is directed to content sources, and it crawls the content in preparation for indexing. You may want to be able to direct the crawler to specific parts of your content.

Content Processing

Content is then processed, using tools in common with initial processing of content by text analytics tools. A text file is created for parsing and indexing, and the content is tokenized, meaning each element is given a unique identifier so that they can be manipulated and located later on. This is also the point at which lemmatization (resolving variant grammatical forms such as tenses and plurals to the same grammatical root) and stemming (resolving different word endings to the same word-stem) occurs. You may also want to remove “stop words” that will create meaningless noise (common operators such as if, but, and, the, a) from the indexing. Your stop words may need to be tuned. Sometimes stop words that are noisy in search indexes provide semantic clues in some rules-driven applications of text analytics.

Indexing

The indexing module creates term indexes from the parsed text. Each term is linked to its source content, and tokenization provides the context in which the search term appears. The pre-processed indexes provide speed in search. Crawling and pre-processing of large-scale content collections can be very slow and occupies significant processing capacity. Hence most indexing consists of incremental updates to the indexes by noticing, crawling and indexing new content whenever it is added. It is very important to plan the configuration and setup of your search engine carefully in advance. Any major changes to the search engine configuration, or to the

way it exploits taxonomy or text analytics may require an expensive and slow complete re-crawl and re-index.

Auto-categorization

Some search engines also offer simple auto-categorization functions. In the absence of sophisticated text analytics tools, these are usually entity recognition based on lookups of reference term lists. These lists can be maintained within the search engine itself, or can be supplied from a taxonomy management system. This is referred to as "known entity extraction via lookup". The entities you are interested in calling out (for example company names or country names) are compiled into lists, and are registered as significant entities if the main search indexes come across these terms or their synonyms. This can provide a very simple filtering capability based on the entity types you are interested in. However, this is a fairly basic technology. Just because a document mentions "Google" does not mean that the document is substantially "about" Google.

Query Processing

Much of the utility of a search engine lies in the sophistication with which it handles queries and query interfaces.

Query interfaces can include the ubiquitous search query box, browsing taxonomy structures where concepts are pre-linked to content through stored searches (requires integration with a taxonomy management module), filtering of search results using metadata elements or taxonomy facets (again requires integration with taxonomy).

Many search engines offer auto-complete or auto-suggest of search queries suggesting common search queries as the user types their query into the search box. Auto-complete can exploit common search queries, existing taxonomy or thesaurus terms (linked to results through stored searches), and/or keywords in context from metadata such as titles of high value content.

The relevance of search results for common queries can be tuned at the back-end by a search manager by altering the rules that calculate the likely relevance of a given piece of content for a given query. Where there is a known "best" piece of content for a given query, this can be promoted to the top of the results page. Relevance tuning can also be automated, by looking at click behaviors between query and clicking on results and promoting content that is frequently clicked on.

Relevance-tuning can exploit what is known about the users and their contexts. Knowing that a user is searching from within a given organizational function enables rules to be written for how queries from those users should be handled, from zoning the search to known target content for that functional group, to handling relevance calculation based on metadata associated with those functions and users, and matching it with the indexed content. Content ranking mechanisms can also be pulled into the mixed.

Search Analytics

Query handling and search tuning require sophistication in the search engine itself as well as a relatively sophisticated search management capability – i.e. the way that search is staffed, administered, configured and constantly tuned in relation to known user needs.

Search analytics is the module that powers this sophistication and tuning. Search engines have very powerful reporting capabilities, tracking how queries are conducted and user interactions with results pages, but as in all complex reporting capabilities, the smarts lie in how the reports are defined and configured. What will you track, how, and how frequently? How will you back up hunches gathered from the reports, and validate them with users? The initial search design strategy based on use cases and content analysis will give you a starting point. This will be refined by the insights you gain by observing user behaviors through the search roll-out and subsequent maintenance.

Taxonomy and Ontology Management

Like search, taxonomy and ontology management requires very robust analysis of how users need to interact with content, and analysis of the content itself to understand how it is described, organized and used. Taxonomy development and management is still best conducted as a work of skilled human design, but it can be substantially enhanced with the appropriate technologies. There are two broad functions within enterprise taxonomy and ontology management systems.

Taxonomy Building and Maintaining

Enterprise taxonomy management systems support the work of building and maintaining taxonomies by providing workflows and metadata around taxonomy concepts. For example, they can house source vocabularies of varying authority and structure that can be used to generate candidate terms for the taxonomy. There are approval workflows to track the approval process. There may be roles-based permissions for complex environments where multiple taxonomy professionals are involved in maintaining the vocabularies. Sources and usages of taxonomy terms can be tracked as attributes of the taxonomy concepts. Scope notes can be added to explicate taxonomy terms, and the revision history of terms and taxonomy facets can be tracked.

Reference Store

The second major function of an enterprise taxonomy management system is to act as a reference store for other systems, supplying taxonomy terms, relationships between terms, synonyms and controlled vocabularies to other systems. It can supply:

- controlled vocabulary terms to content management systems to support manual tagging of content
- terms and relationships to support auto-categorization via known entity lookup, whether applied within a search engine or a text analytics engine;

- synonyms and relationships to search so that it can improve relevance and precision of results, and support expansion of search
- taxonomy facets to search to support filtering of search results.

Because they are built to manipulate defined relationships between concepts and their attributes, taxonomy and ontology management systems can also act as brokers to other Linked Data or Linked Open Data sources, to call authority terms and relationships from elsewhere, to support search and/or to support enrichment of content via text analytics.

Finally, enterprise taxonomy and ontology management systems are built to act as taxonomy/ontology hubs catering to multiple systems and audiences. They are capable of interacting differently with different platforms and audiences. For example, different versions of the same core taxonomy can be presented differently to different audiences, depending on their needs. Different vocabularies and taxonomy services can be supplied to different systems based on their content and uses.

Text Analytics

Text analytics refers to a quite diverse family of computer-assisted techniques for assigning meaning to content. Different modules can be used to serve different purposes.

Content Processing

There are some basic content processing functions that text analytics has in common with enterprise search: preparation of text for parsing, tokenization, lemmatization, stemming, stop words.

Syntactic and Semantic Pre-Processing

Several functional applications of text analytics depend on two foundational processing techniques.

The first, syntactic analysis, analyzes the language structure of text grammatically, using very large open source training sets. Sentence splitters break up sentences into parts of speech, distinguishing nouns and noun phrases from verbs and operators, and so on, and syntactic analyzers create a logical representation of the sentences.

Semantic analysis attempts to infer meaning from analyzed texts. Like syntactic analysis it depends on reference training sets. It enables computers to infer meaningful assertions and facts from bodies of text, and underpins the base technologies behind machine translation. Syntactic analysis is based on known rules of grammar. Semantic analysis has a much more challenging task, which is to infer human-understandable meanings. Not all grammatically correct sentences are meaningful to humans – "Colorless green ideas sleep furiously" is a famous example

presented by Noam Chomsky of a syntactically correct sentence that is semantically incoherent.

Syntactic and semantic techniques underpin the identification of "entities" within content. In the context of text analytics, an "entity" can be a person, a thing, a place, a time, a topic. It is any distinct concept that can be identified from text. Identification and extraction of entities can be done most simply by simply looking up reference lists, as in auto-categorization of known entities provided by some search engines.

However, with syntactic and semantic analysis text analytics can also identify unknown (unpredicted) entities for which you don't yet have examples in your lists. For example in English proper names are capitalized, and personal names have known forms (that also vary by culture). Hence candidates for persons mentioned in text could be picked up by a rule that says "look for one or more capitalized subjects or objects of sentences, and then look for lemmatized verbs from the following list 'say'… 'reply'… etc".

Such rules are based on common usage as represented in their training sets and are typically provided as standard. These rules will typically need to be tested and refined based on local usage. For example, we found that one standard tool did not pick up country of origin in content from an immigration and customs authority, because it was using simple term-phrase lookup. However, customs officers at checkpoints habitually used the convention "[Country]-registered vehicle". The hyphenation and extension of the term disrupted the simple lookup function. Notice that the solution to this problem is a combination of techniques, likely supplementing known entity lookup with syntactically driven entity extraction rules. The practice of using different techniques in combination of characteristic of text analytics approaches.

With these foundational capabilities, there are many specialized applications of text analytics, not all of which will be suitable to every purpose.

Auto-categorization

The ability to identify both known and unknown entities/concepts is a significant improvement in auto-categorization capabilities. Lookup lists for known entities/concepts can be supported by a taxonomy management system, and strengthened through the supply of synonyms.

Conversely, through the judicious use of rules, text analytics can discover entities/concepts mentioned in text that are not yet in the taxonomy or controlled vocabularies, and can be considered as candidate terms.

Text Mining

Text mining refers to a number of techniques for analyzing text based on statistical algorithms. In very broad terms, large bodies of text are submitted to iterative

operations using algorithms to “sort” the text into clusters or “buckets” that are internally statistically similar, and statistically dissimilar to the other clusters or buckets.

Statistical techniques are also used to determine the words or word-phrases that are most likely to uniquely represent their bucket compared to other buckets. These word strings are often called “topic models”. The likelihood of specific terms occurring in proximity to each other can also be calculated.

Text mining is sometimes touted as a fully-automated alternative to describing and tagging content. In practice however, the clustering techniques use fuzzy algorithms, and the end “described” state is highly influenced by which clustering pathways are taken at the beginning of the operation. Text mining techniques suffer from non-reproducibility (the same series of operations on the same content can produce different results on different occasions) and from non-comprehensibility – the topic models generated often do not match how users themselves describe the content. So in practice, applications of text mining are heavily tuned by their human operators, and it is not always transparent how a given set of results are achieved.

Text mining also suffers from endogeneity – meaning that the tags that are extracted to describe similar buckets of content are wholly derived from that content set (or when training sets are used for comparison, from the target content + the training set). Endogeneity is a problem when you want to survey diverse collections of content and map meanings consistently across them. In that case, a taxonomy or ontology that crosses those content sets and audiences provides an exogenous reference point that can broker meanings across different content sets and user audiences.

However, text mining can be a powerful tool used in combination with search and taxonomy management. Its statistical techniques to cluster based on similarity can propose new categorizations and potentially relevant relationships between topics to the taxonomy manager. Its ability to extract topic models can be useful in automated summarization techniques.

Semantic Analysis

We have already discussed some of the challenges of complexity facing semantic analysis, not least the heavy influence of context and culture (factors external to the content) on meaning. In practical text analytics applications, semantic analysis relies on a number of techniques and not just computational semantic analysis. For example, one of the main applications of semantic analysis is to extract assertions or facts from complex documentation. If semantic analysis has access to ontologies via a taxonomy management system it can then enrich its inferences based on known and validated relationships between entities.

Sentiment Analysis

Sentiment analysis is a very specialized application of text analytics, often used in marketing, social media analysis, and qualitative feedback. It looks at the semantics of positive, negative and neutral expressions, in association with known (extracted) entities. Again it is highly contextual, and subject to over-simplification – for example, sarcastic language can be superficially positive, but the linguistic markers of sarcasm are often not very obvious. As with most text analytics techniques, the rules and algorithms need to be tested, validated and constantly supervised for accuracy.

Summarization

Automatic summarization often exploits known document or data structures (it knows where to look for key content), and it may exploit syntactic analysis (e.g. removing extraneous "padding" language such as adjectives and adverbs). It may use semantic analysis to extract key facts from documents.

How the Pieces of the Stack Work Together

Let's bring this back together to get a better understanding of how the pillars of the stack work together. Figure A2 illustrates the high level view of how the different pillars of the search and discovery technology stack interact and work together to produce superior results for organisations. The focus for how the stack is deployed, and which particular technology components are used, always springs from user analysis, and an understanding of the queries that users make, against an understanding of the target content, how it is structured, how and why it is produced, how it is currently being used, and how it could potentially be used, if the search and discovery functionality could be improved. In digital environments, "target content" can be any combination of documents, web content, data, dashboards, and people profiles.

The configuration of the elements of the Search and Discovery stack springs from an understanding of user needs against an understanding of target content.

Determining how to use the pillars of the stack sits within a larger diagnostic and decision-making framework.

1. What is the business problem (or set of business problems) you are trying to solve? The answer to this question can often come through a knowledge management diagnostics activity such as a knowledge audit, combined with an understanding of the organisation's business direction and strategy.
2. Who are the key user communities in relation to the business problem, and how will you understand their working needs and opportunities for improvement? (This can also come from a knowledge audit).
3. What is the target content (the knowledge resources that your target user communities needs to work with more effectively), where is it, how is it structured and used currently? (Content analysis and modeling can help).

4. What tools from within the search and discovery stack can help to address the proposed solution? (The focus of this paper)
5. What needs to change on the behavioural, process and governance levels for the new solution to work? (For sustainable implementation planning, check out *The Knowledge Manager's Handbook* (Milton and Lambe, 2016) from a knowledge management perspective or Maish Nichani's series of articles at http://olasearch.com/articles from a search design perspective).

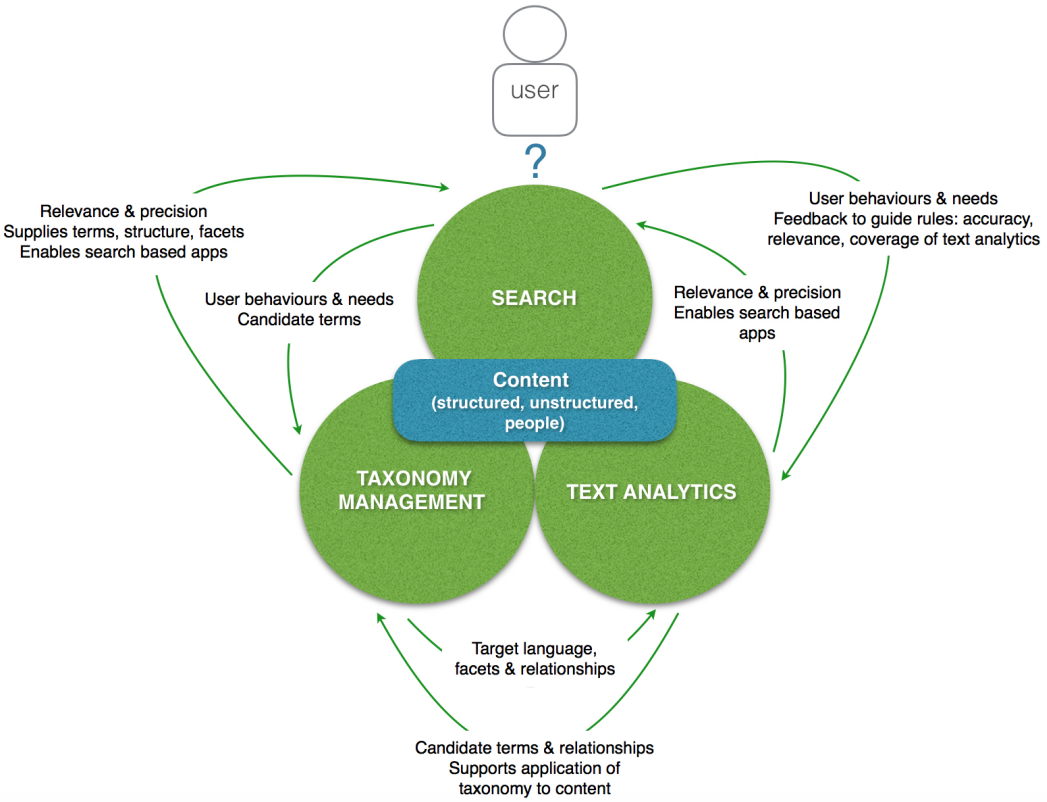

Figure A2 Interactions between taxonomy management, search and text analytics

Search and Taxonomy

Search is the user-facing component of the stack. It mediates content to users. Taxonomy management supplies known salient concepts and relationships to search, and it tells search which query terms are synonyms of the same taxonomy concepts. By "salient" we mean that the concepts are front of mind, action-related concepts in the heads of users as they perform information and knowledge seeking activities to complete work tasks. Taxonomies make search smarter by telling it which concepts are important to users, how they are related to other concepts, whether through hierarchical or other relationships. Hence search can exploit taxonomies to help users to refine or expand their search, to follow meaningful pathways and explore content. Taxonomies and ontologies enhance the relevance and precision of search for users.

Because search is user facing, it gathers a lot of data about user needs and behaviors. Frequency analysis of search queries tells the taxonomist a lot about how users think about their content and their search requirements. Click through activity

on search results pages also provides evidence about user perceptions of what may or may not be useful. Hence search analytics provides a feedback loop to help the taxonomy manager constantly refine and improve the taxonomy to user needs. This is a positive feedback loop, where taxonomy provides relevance and precision to search, and search provides a constant flow of real time large-scale evidence on what is actually relevant and useful to users at the present time. Search analytics can also pick up emerging terms and concepts that are important to users, and propose them as candidate terms for taxonomies and their supporting thesauri.
Without taxonomy management, search is relatively dumb, whether in terms of what it can index, how it can help users make their queries, and how it can help users act upon their results. Without search, taxonomy management lacks a convincing medium to demonstrate its value, and it lacks a continuous flow of evidence to maintain its relevance and usefulness to users.

Search and Text Analytics

Text mining can propose new concepts and new associations between content items for the search manager to build into query processing and relevancy tuning. Semantic analysis and summarization techniques can be used to provide "smart" summaries of content items to be previewed in search snippets and in search results pages, making it easier for users to assess whether or not the content is relevant to their need, before they open the content item. Text analytics can extend the way in which content can be pulled together in specialized search based applications.

All three pillars of the stack (search, taxonomy and text analytics) depend on constant tuning of the infrastructure to deliver good search and discovery results. Text analytics in particular depends on rules to guide how the content is processed and tagged. These rules need maintenance and tuning. This tuning is necessary because of the constant ongoing changes in content, context, priorities, users and needs. Search analytics provide a large scale, real-time window into user needs and behaviors, and so, as with taxonomy management, it provides the evidence base needed to keep the text analytics stack tuned to deliver accurate and useful results for users.

When there are large and complex knowledge bases, without text analytics, search finds it difficult to pick out salient concepts related to content. As with taxonomy management, text analytics depends on deep user and content knowledge to know how to configure and tune its operations. Without search, text analytics finds it hard to do this unless the team is prepared to invest in constant (and expensive) user research and testing.

Text Analytics and Taxonomy Management

Taxonomies are designed, evidence-based artifacts. As such, they consistently lag the pace of change. They are optimized to exploit known and documented concepts. Text analytics techniques such as auto-categorization and text mining can supply candidate terms and relationships to the taxonomy manager without the need for

comprehensive re-surveying of content and user needs. They can pick up emerging concepts in the content, just as search analytics picks up and proposes emerging concepts evidenced in search queries. So text analytics can act as a powerful taxonomy enrichment tool.

Taxonomies and ontologies can guide and strengthen the application of text analytics techniques such as auto-categorization (known entity extraction through lookup), fact extraction through semantic analysis, and sentiment analysis (by helping the sentiment engine resolve different synonyms to the same entity being referenced in the content).

When there are large and complex knowledge bases, without text analytics, taxonomy tags cannot easily and consistently be applied to large, diverse bodies of content. Without taxonomy management, text analytics struggles to describe content in terms that make sense to key users and in the context of user activities. Machine tagging techniques alone do not describe content in terms that humans intuitively understand, without some form of human-curated filter. Taxonomy management provides this.

Implications

For too long organisations have been listening to the booming voice giving magical reassurances from behind the curtain, and have neglected their own capabilities, and the toolsets that exist in the open source arena (that commercial products also exploit). As the world becomes more complex, solutions to search and discovery problems will become more diverse. We need to become better journeymen, more knowledgeable about the capabilities different toolsets that are available, so that we can choose the right tools for any given problem. Some of them will be commercial tools, acquired for their excellence in a specific set of functions. Some of them will be open source tools, acquired and tuned to solve specific kinds of problem. The age of single source “do it all” commercial platforms is over.

This paper has proposed that knowledge of the search and discovery technology stack is an important capability to acquire. It needs to be at least adequate to the job of evaluating needs and possibilities, and of asking vendors searching questions about how their technology works, what it is good at, and what its limitations are. In some cases, organisations will want to internalize a deeper technical capability in working with a number of sets of tools.

There are other important search and discovery related capabilities that enterprises need to acquire beyond a knowledge of the search and discovery technology stack – for example, how user needs and content can be analysed to determine priorities and opportunities, how content can be managed and enriched to provide optimal user experiences, and how business problems can be transformed through design processes into effective solutions for business needs, problems and opportunities.

I would like to acknowledge the following colleagues who have provided valuable insights and advice on this paper: Dave Clarke, Agnes Molnar, Maish Nichani and Tom Reamy.

Patrick Lambe October 2016-January 2017