IFLA WLIC 2016 COLUMBUS

Submitted on: 25.10.2017

Sharing Practices and Actions for Making Best Use of Organizational Knowledge in Libraries
August 12, 2016
Langsam Library, University of Cincinnati
Cincinnati, OH, USA

Building Knowledge Strategies with Search Data

Drew Wiberg
Department of Information Technology, Eppstein Uhen Architects, Milwaukee, WI, USA.
E-mail address: dreww@eua.com

Abstract:

This paper outlines a toolkit to be used by a knowledge manager or corporate librarian charged with a knowledge management initiative. The toolkit is built of four strategies for connecting a corporate community with existing and new information resources that use a single dataset, the community's intranet search query logs, as the foundation. In order, the strategies allow a knowledge manager to improve intranet search engine performance, design a more relevant and efficient information architecture for an existing intranet, create new content that directly fills identified knowledge gaps, and inform the curriculum of an existing corporate university program. The visualizations used to perform the analysis in this paper are built using Tableau and Gephi. Assumptions are made that the reader has a small degree of familiarity with these tools (or BI tools similar in design) in order to maintain efficiency of prose in the methodologies. Another major tool used is Knowledge Architecture's Synthesis platform, particularly the Synthesis Search Optimization dashboard, which at the time of publication, is still in a beta release. Knowledge management takes many forms and this paper does not dive deeply into the theory on why these tools are effective. Assumptions are made that they are applicable to existing KM initiatives or are adaptable to the needs of the reader.

Keywords: Knowledge Management, Search Queries, Intranet, Content Creation, Internal Education

1.1 INTRODUCTION

This approach to Knowledge Management was born out of intense anxiety. After a year and a half of work as a knowledge manager for the mid-sized architecture firm, Eppstein Uhen Architects of Milwaukee, WI, and only one year after completing my Masters of Library and Information Science degree at the University of Wisconsin - Milwaukee, I was asked by Chris Parsons, the founder of the software company Knowledge Architecture, if I would present an overview of my KM strategy at their annual knowledge management conference, KA Connect.

When Chris first asked me if I'd like to present my work in KM at KA Connect, my gut reaction was to run away. My brain reeled at the thought of being on stage in front of that many people. I'd only been on stage in front of an audience the size of KA Connect's once, when I was eight years old, for my grade school's production of *The Legend of Sleepy Hollow*.

I had the indistinct role of Ichabod Crane's fiancé's father, was in one scene - when Ichabod came to a dinner party in his honour - and I had one line:

"Have a seat Ichabod!" that I was to deliver while motioning to a chair at the dinner table that was to be set centre stage. I took my role very seriously for an eight year old, I practiced at home, attended all of the rehearsals; I was very engaged with the theatrical process. On opening night, in front of a full gymnasium of parents, teachers and older kids, I walked onto stage at my cue, stood on my mark, and delivered my line:

"Have a seat, Ichabod!" and motioned to an empty place on stage; someone had forgotten to place the chair for Ichabod to sit it. The mistake put me into a terror and I quickly made an exit stage right, to the delight of the audience.

This early experience kept looping through my head over and over as I tried to think of a way to describe my approach to knowledge management at my firm. In this anxiety-induced fever dream, I came to the final realization that the chair and table could be a workable metaphor for the KM program I was charged with implementing.

The table, in this metaphor, is where all of a community's conversations happen and the chair is a Knowledge Manager's strategy. If the chair is missing, if I don't have a clear strategy, then there will be no way for me to help capture my firm's conversations or connect community members with the knowledge resources that they need.

Trained as an information scientist, I am comfortable wrestling with data, especially the data stored in search engine logs. Below is a representation of over 5000 unique search terms collected from 14000+ search queries put to our intranet's knowledge base at EUA over the past 9 months.

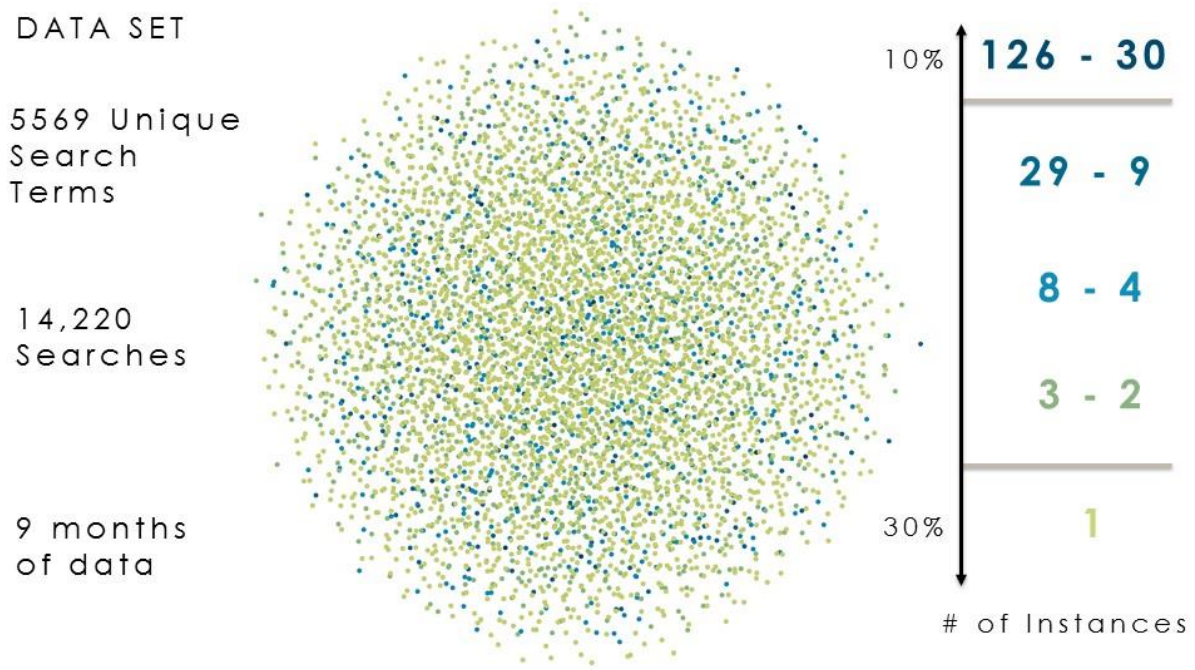

Figure 1: EUA's Search Dataset

There is strong evidence in the body of information search and retrieval research that the top 10% of search queries in a dataset from a homogenous group will reflect or accurately represent the entire dataset (Quesenbery at all 2008, p. 3). By concentrating on this subset, analysis becomes more efficient; the first two legs of the chair, of the KM strategy, will use this subset of the data.

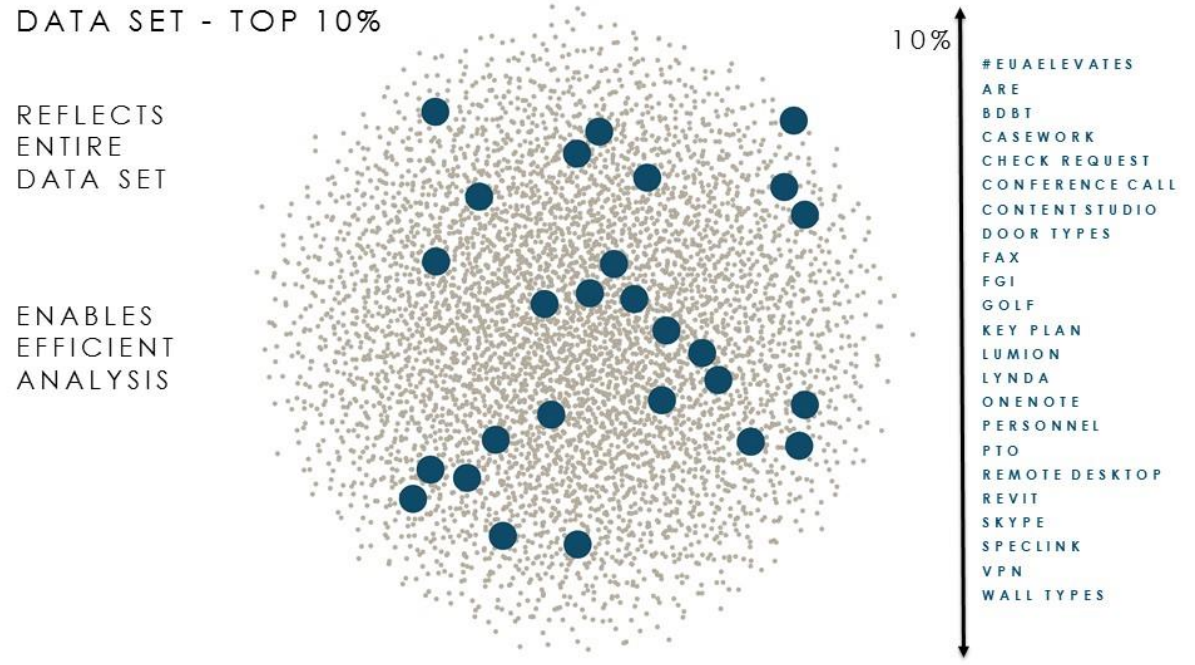

Figure 2: Top 10% of EUA's Search Queries

2.1 IMPROVE SEARCH

Improving the performance of your intranet's search engine is one of the highest impact methods of connecting others to the knowledge base, it also has one of the lowest resource costs of this KM strategy.

Information seekers are plagued with non-relevant search results, this is true in the wild, open internet as well as on corporate intranets. Search technology has not progressed quite to the point where it can intuit a seeker's search interests from an initial query. This is primarily due to the diversity of natural language (Shi 2007, p. 1), but another issue that contributes to non-relevant search results is the lack of knowledge by the seeker in how the search software matches her query with the documents it produces. Search engines are overwhelmingly a 'black box' as far as the information seeker is concerned. If the user could participate in the internal retrieval process somehow, relevance of search results would naturally jump. The normal information seeker does not have the access or training to engage with this process (Zhang 2007, p. 2).

It is possible, though, for an information seeker to influence this process simply by using the search engine in a normal way, through the behind-the-scenes work of a knowledge manager and the efficient analysis of those search query logs. A knowledge manager trained in seeing the patterns in search data can help identify, for instance, top performing and underperforming terms in a dataset. (Quesenbery at all 2008, p. 3). After identifying underperforming terms, it is possible to dive deeper into individual search sessions, where the information seeker, through normal behaviour, can influence the relevance of search results.

As mentioned, search queries in a large dataset, especially one culled from a homogenous group of users such as those using a corporate intranet, follow predictable patterns. One of these patterns are that the top 10% of search terms will normally reflect the bottom 90% of the same set. When looking at the subset of the top 10% of terms it is possible to break that subset into top performing and underperforming terms and follow those underperforming terms to their network of semantically related siblings in the bottom 90%. Through working to deliver good results for the top 10%, those improvements can be felt for all information seekers on the system.

There are a number of methods that a knowledge manager can use to perform this analysis. The first method we will discuss is visual analysis. Visual analysis can help identify patterns in a dataset quickly (Stenmark 2008, p. 2233). These patterns can reveal usage or session patterns of users or self-organizing patterns that exist through related terms (Quesenbery at all 2008, p. 9). This type of visualization is derived from what is termed a graph, where graph is defined as a representation of connections between terms, or nodes, based on preselected criteria (Hu and Shi 2015, p. 115). Graph networks can become unwieldy quickly as the original dataset gets larger, so this form of analysis is only efficient if the dataset can be parsed into smaller chunks, the top 10% subset for instance. When looking at the full dataset at once, another approach is required. Through the weighting of terms based on an aggregation of performance indicators and assigning each term a score or some other value based on this weight, it becomes easier to identify how individual terms perform overall, even across the entire dataset.

The search session is rich in insights that can benefit search result relevance. Search goals, complexity in search strategy, and the difficulty of the search task are all data points that can

be pulled from search sessions. The search goal is the primary data point that can be identified from the search session. This is indicated not just by the first search term used but by semantically related, proceeding terms in the same session. Complexity of search strategy can be seen in diversity of related terms, that is, if the proceeding terms are just small variations of spelling or are closely related synonyms, the complexity can be considered to be relatively low. On the other side of the scale, if the proceeding terms in a multi-query session are distantly related, showing an information seeker driving for relevant results using terminology from different disciplines or Boolean logic operators, the strategy can be considered to be more complex and weighted accordingly.

Ultimately, a search query session does not record the real reasons the seeker engaged with the engine, which is a known limitation (Jansen 2006, p. 424). The safest assumption a knowledge manager can make is that the information seeker wanted precision and relevance in the results returned (Shi 2007, p. 10). A knowledge manager for an organization has the added benefit of context and of knowing many of the information seekers on a professional or personal level (Athukorala et al 2015, p. 4). She can often identify not only strategies to improve search results through looking at small scale network graphs, large scale aggregated scoring and search sessions, but her contextual knowledge of the microcommunities in her company can give her insights on where and who to go to in order to extract new content that can be then added to the knowledge base. Connecting members of these communities with the right information at the right time and identifying areas that need new content should be at the core of a knowledge management initiative.

2.2 IMPROVE SEARCH - METHODOLOGY

A number of tools are used to collect and analyse the data presented in this paper.

The first being the business intelligence software, Tableau:

http://get.tableau.com

Tableau is self-described as “business intelligence software that helps people see and understand their data.” There are a number of BI tools available to the public now and their range of difficulty extends from intuitive to involved. Tableau is, in the opinion this researcher, one of the more powerful of the available tools, but the learning curve is understandably steeper as compared to others.

The second is the social intranet platform Synthesis, from Knowledge Architecture.

http://knowledge-architecture.com/synthesis.php

Knowledge Architecture [KA] is a software company from San Francisco that specializes in serving the Architecture / Engineering / Construction [AEC] industries. KA is self-described as a company that “build[s] intranets for architects and engineers. In addition to their software development they hold knowledge management workshops, host knowledge management communities of practice, and host the aforementioned annual conference, KA Connect.

Synthesis is built to integrate with Microsoft SharePoint:

https://products.office.com/en-US/sharepoint

and EUA, at the time of this writing, uses the SharePoint Foundation 2013 product as a platform to host KA Synthesis.

The graph network visualizations for this paper were built using Microsoft Excel:

https://products.office.com/en-us/excel

and the open source graph visualization program, Gephi:

https://gephi.org/

Synthesis Search does not rely on the out-of-the-box SharePoint search, instead employing the search engine technology, Elasticsearch:

https://www.elastic.co/products/elasticsearch

The initial dataset was created by connecting Tableau to the search query logs hosted locally via SharePoint using the following method:

1) Connect Tableau to the appropriate table following the below path:

   Search_Service_Application_LinksDB
   - ➔ Tables
   - ➔ File Tables
   - ➔ Dbo.MSSQLogPageImpressionQuery

2) This table stores faceted search queries, which means that for every query imputed by a community member, the result is 6 to 8 entries in the log. To cull the redundant entries a wildcard filter was created on the query strings to only show entries that contain the following:

   *SPContentType=Link

3) Results were then further filtered to cull the machine-readable text and only display the actual user query using the 'Create Calculated Field' tool under the 'Analysis' tab in Tableau. Three calculated fields were created, nesting each in the other until the desired result was achieved.

   a. Start at 3: MID ([Query String], 3) – result: string is displayed beginning at the third position
   b. Truncate: REPLACE ([Start at 3], " ")) AND (-owsURL:connect+eua+com) AND SPContentType=Link", " ")
   c. Upper Case: UPPER ([Truncate])

The display in Tableau was built using the 'horizontal bars' option, this visualization is basic but useful. It displays the query term and the number of instances the term has in the log. It has the added benefit of being able to effectively copy / paste into an Excel spreadsheet in a clean manner.

Figure 3: Snippet of full dataset as visualized using Tableau

In addition to using Tableau to scrape the entirety of the SharePoint search logs for the purposes of these exercises, EUA was in the beta test group for Synthesis 5.2: Search Optimization and the Synthesis Search Optimization [SSO] dashboard was leveraged heavily to investigate the dataset.

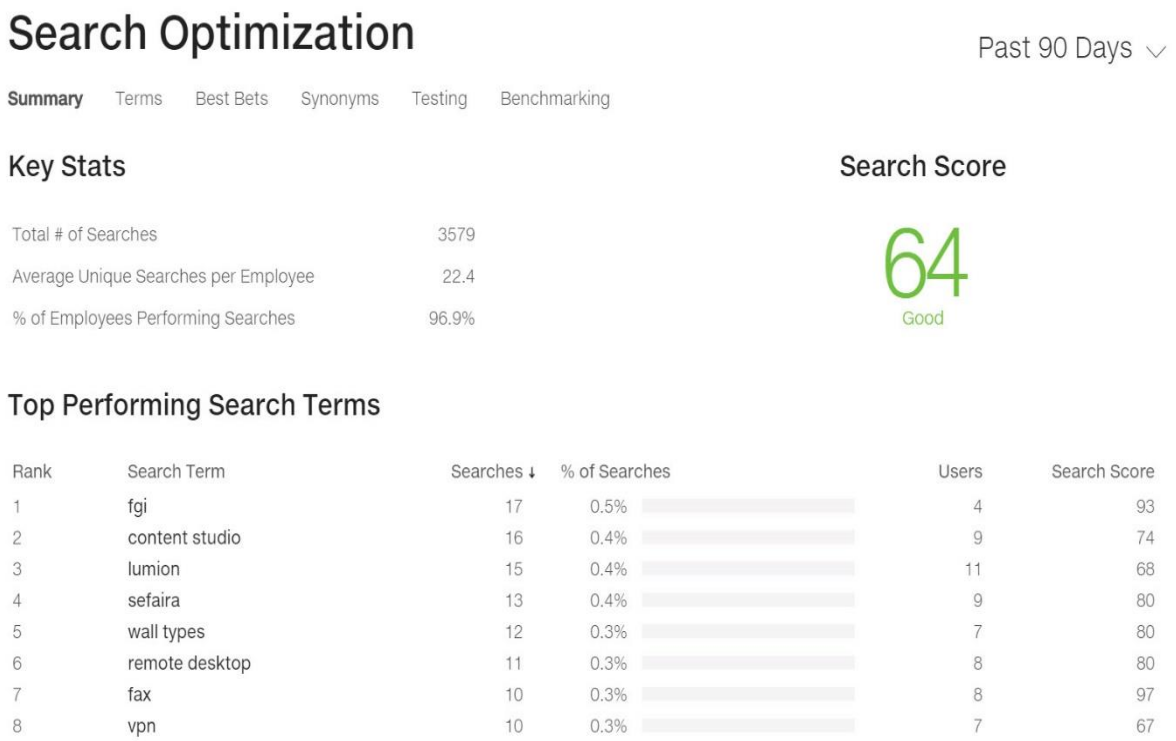

Figure 4: Summary view of Synthesis Search Optimization [SSO] dashboard

As stated, Synthesis Search employs Elasticsearch technology on the back-end and ReactJS (http://facebook.github.io/react/) to display the results. The above illustration shows a sample of top performing terms. Top Performing and Underperforming terms are displayed with a score that is weighted on the number and type of actions taken in a session, where a score of 100 denotes a session where a query was entered and a relevant result was found and clicked on by the information seeker, as is shown in the below image.

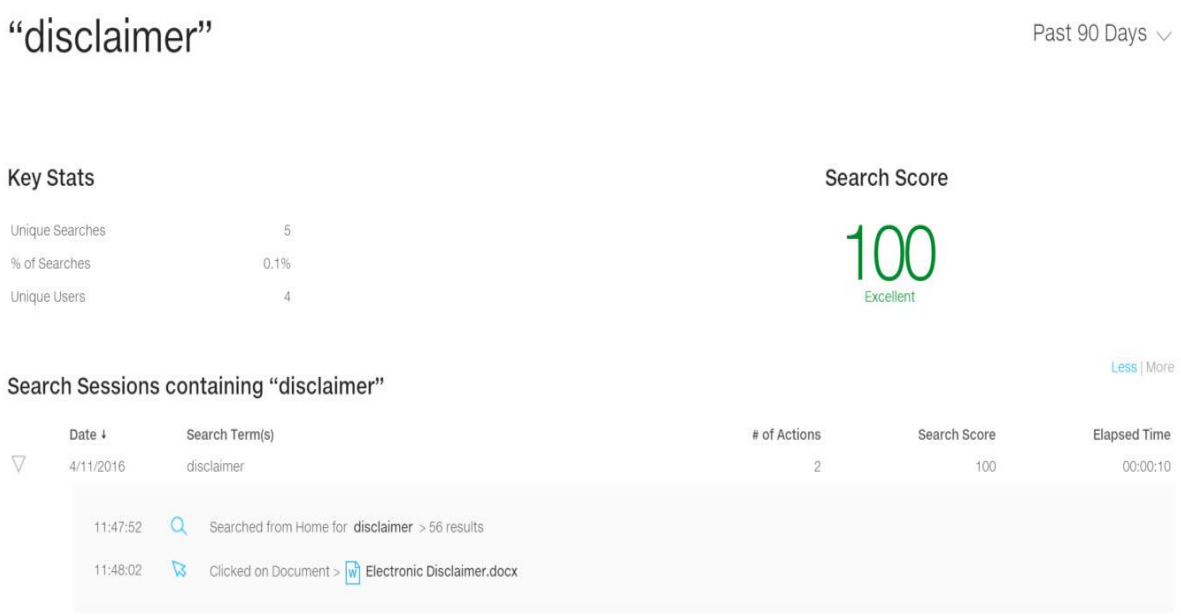

Figure 5: View of single term with a search score of 100

A score of 0 denotes the opposite, where a query was entered and no relevant result was found by the information seeker as indicated by the lack of the session ending on a significant information resource. These scores are used as guideposts for efficient analysis of the dataset.

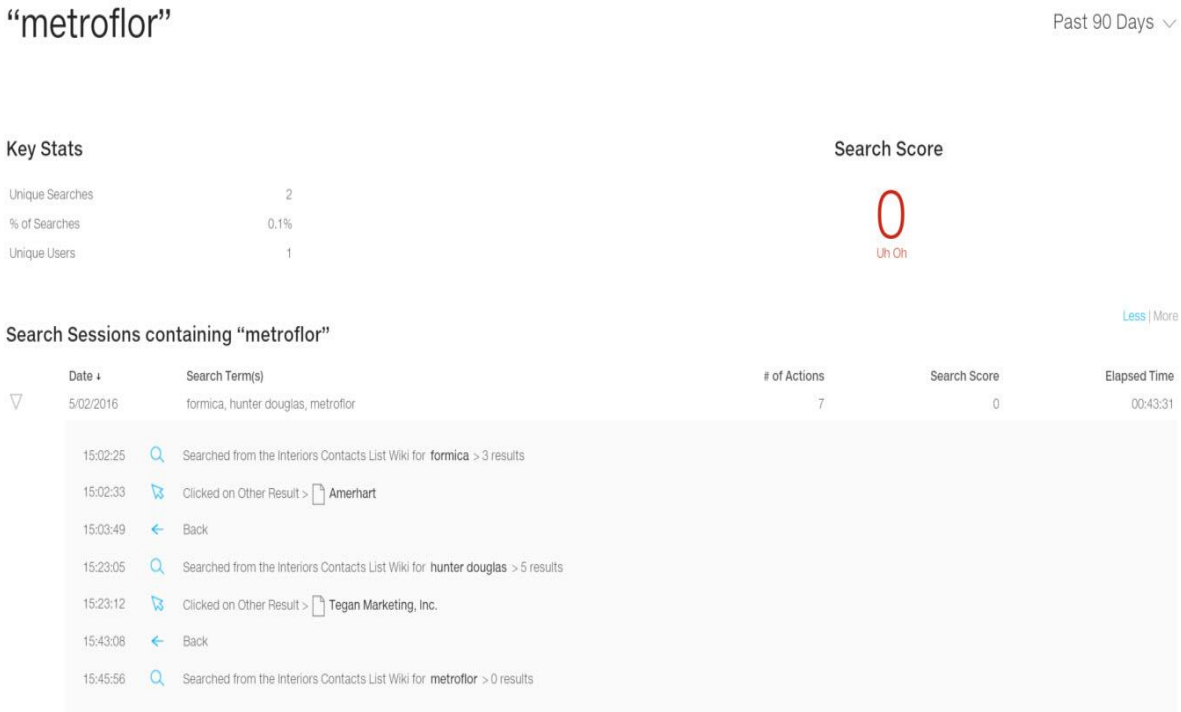

Figure 6: View of single term with a search score of 0

2.3 IMPROVE SEARCH - RESULTS

To utilize these tools to improve search functionality, the synonyms panel of Synthesis Search was used after underperforming terms were identified via a cross reference against the top 10% of queries in the dataset using the previously shown visualization from Tableau.

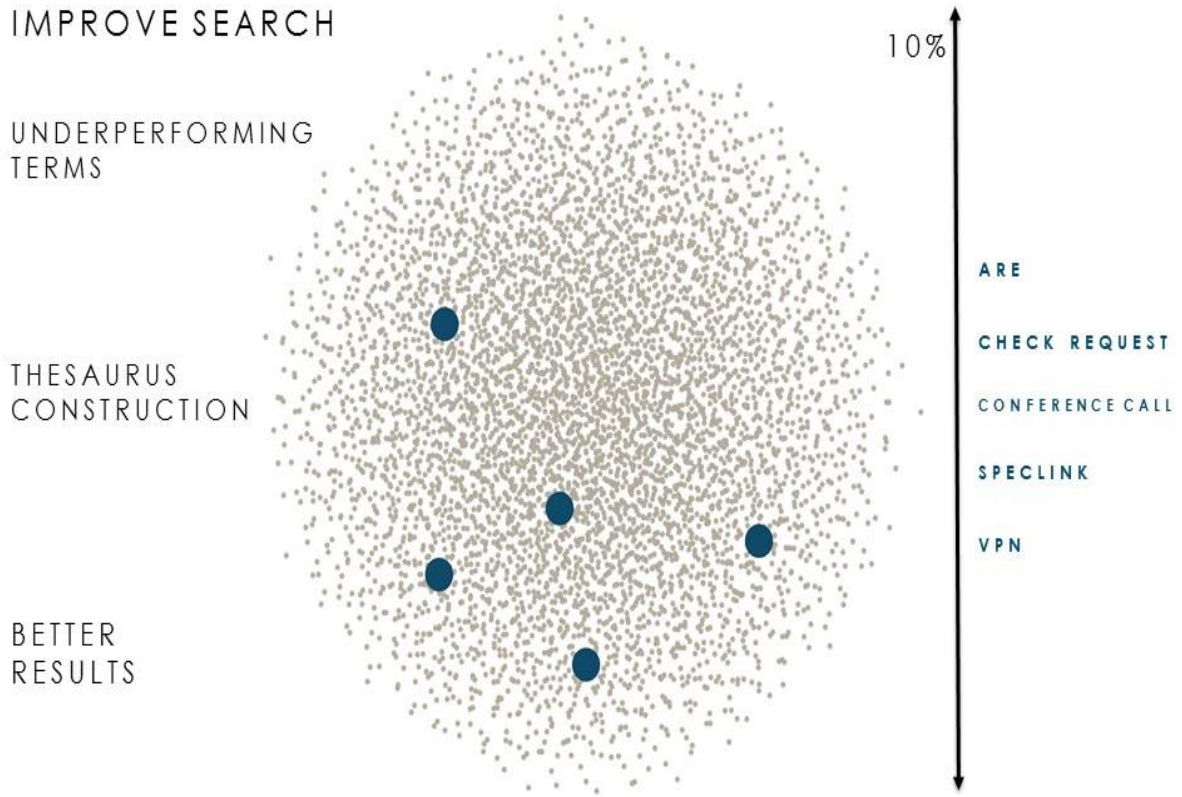

Figure 7: Underperforming terms in the top 10% subset

ARE [Architect Registration Exam] is a good example of a term that is searched for a lot on EUA's intranet, but at the time of analysis, users weren't finding relevant resources.

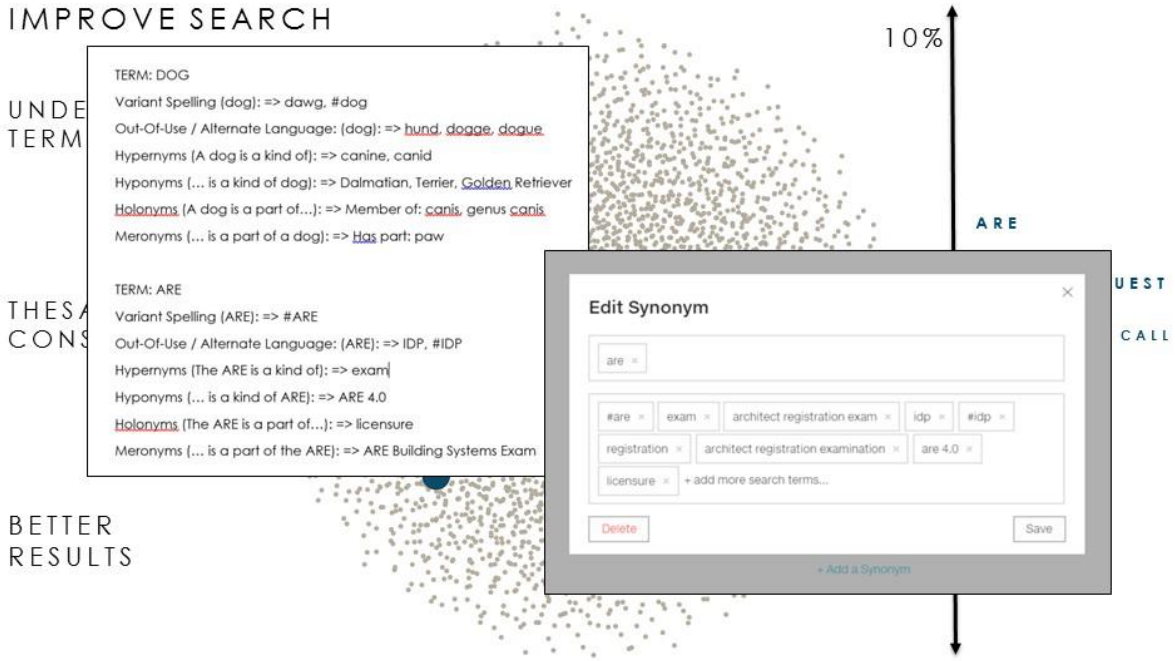

Figure 8: Basic Thesaurus Construction guide and the Synonym Edit pane in SSO

We can help out here through the very traditional information science technique of thesaurus construction. Through adding synonyms to the term 'ARE' in the search optimization panel the reach of the query is broadened. Anytime an information seeker searches for one of the synonyms chosen for that term, they see results prioritized for ‘ARE.’ Through this search optimization process, the preferred results can be added as a 'Best Bet' result for their search. Even in this age of technology, you can't beat a human's ability to see patterns and to shape how knowledge is connected.

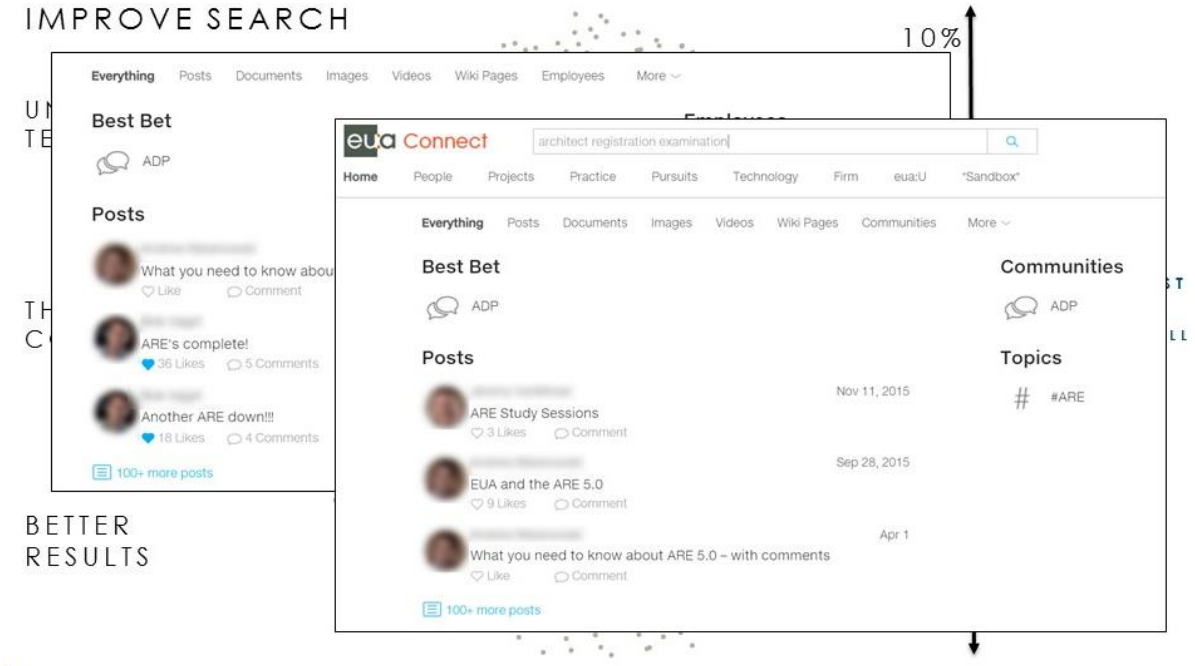

Figure 9: Synthesis Search results exhibiting a pre-determined ‘Best Bet

Now this is a useful trick, sure, but not an entire KM strategy to be sure, so far we only have a one-legged chair. Let's keep building.

3.1 DESIGN NAVIGATION

Another way to employ this raw data to the purposes of KM is to use it to inform the design of the intranet, its information architecture, making it more useful for a community of users. Highly rated search terms in a dataset can suggest places where the information architecture could be improved. For example, a large number of searches for a particular web page could suggest that the existing design is failing users and adding a link or featuring that page in a prominent way will improve their experience (Quesenbery at all 2008, p. 5). This behaviour is easily differentiated from the 'normal' information retrieval process if the analyst has sufficient information in her dataset or context about the community using the search engine. For enterprise search on an intranet knowledge base, both are usually obtainable. Information seekers should exhibit behaviour that suggest they are pushing the engine to provide them with a set of relevant results, and not one result in particular (Zhang 2007, p. 2).

For the purposes of this paper it is accepted that those search queries in the peak of a dataset represent or are strongly indicative of the tail of the same set (Jarrett et al 2009, p. 735). The method of analysis can be much quicker by concentrating on that top tier in this instance as well as the last. Analysis can be made more efficient if the assumption is made that navigational searches can be identified even quicker by understanding what they are not. Navigational searches are not exploratory or informational searches. Using the search engine to browse information resources has a stark contrast in search session logs from navigational searches, For instance, they have a higher mean of cumulative clicks. Exploratory searches show seekers spending more time living with their results, sometimes extending the length of the session into an hour or more as they read through or preview the results, returning again and again to scroll deeper and find more information (Athukorala et al 2015, p. 15). Informational, or knowledge acquisition, searches have more mean click-throughs than exploratory searches (Athukorala et al 2015, p. 12) but still fall under the number of click-throughs for the navigational search. Navigational searches will typically exhibit one query and one click-through on a document that is a shared destination across multiple unique information seekers. Further, they will normally not exhibit any 'learning loop' behaviour. A learning loop in a search sessions shows the information seeker attempting a query and then modifying that query based on the results that she is presented (Pirolli and Card 2005, p. 2). When performing a data analysis with the goal of identifying navigational searches, one should be looking at query length, maximum scroll depth, and task completion time. A navigational search will have a short query that contains no advanced logic operators, will have almost no scroll depth, and will have a task completion time under 5 or 10 seconds (Athukorala et al 2015, p. Abstract).

Using search data to identify navigational searches is a form of co-design. Co-design traditionally taps into the creativity of those tasked with design and people that are not trained in design (Sanders and Stappers 2008, p. 2). Scraping search data for user behaviour is a method that allows a designer to view information seeking behaviour, to interact with her co-designers, with no impact on the resources or time of the second party. Co-design of information architecture using navigational searches does, by definition, need to happen post-launch of an intranet since a mature dataset of search queries is needed to implement this stage of the strategy. Using the insights from this analysis can have positive, long-range, scalable impact on the success of the new page and the site as a whole, through the application of 'participation at the moment of [the] decision' of the information seeker. A co-designed intranet page gives

the user the position of expert. Utilizing search data to infer the direction that design takes is a manifestation of knowledge management. When a knowledge manager can leverage the tacit knowledge of the intranet community member, using their knowledge of what information they know they need to get to quickly but may or may not be able to express when asked, then the users become knowledge experts of their own experience, generating insights for the knowledge manager at no additional resource cost to themselves (Sanders and Stappers 2008, p. 8).

Utilizing search data to inform intranet design is a superior method to the standard user experience [UX] analytical methods. Examples of traditional methods include volume measures that show number of visits to each page, activity volume that shows the number of log-ins, number of active users, proportional activity, which juxtaposes activity volume against number of total members, average time spent, average number of posts on the social newsfeed, average depth of visit, and so on. Weighted activity metrics measure proportional activity weighted for relevance of specific community needs (James et al 2014, p. 34). An analysis of search data for navigational searches captures much of the same data above in addition to data on community activity that would be missed by the above methods. The search engine is fast becoming the normal first navigation method for most people on intranets and in the wild. Information seekers are leveraging a new postmodern level of information literacy and adapting to poorly designed intranets, or poorly implemented content management systems, and finding direct routes to the information that satisfies their need. The days of browsing a site, especially one as potentially complex as a corporate knowledge base, are nearing an end. Only the clearest, most relevant design and information architecture will survive and find use as the workforce transitions to more and more digital natives. Leveraging this strategy can put the knowledge manager at an advantage over her peers. She will have, at her fingertips, hard data that drives decisions on how to present and organize her knowledge base so that her community members get the information they seek, and at no extra resources costs to themselves.

3.2 DESIGN NAVIGATION - METHODOLOGY

To arrive at the below results, the SSO was used to identify the best performing terms in the top 10% of the dataset, as it was visualized using Tableau. This visualization was pre-built for our previous work in improving search and explained in the last methodology section. Once navigational search behaviour is identified than that data can be applied to making decision about improving the efficiency of the intranet's design.

3.3 DESIGN NAVIGATION - RESULTS

Navigational Search Behaviour has been previously characterized as a corpus of very quick searches and successful sessions that end in a common destination.

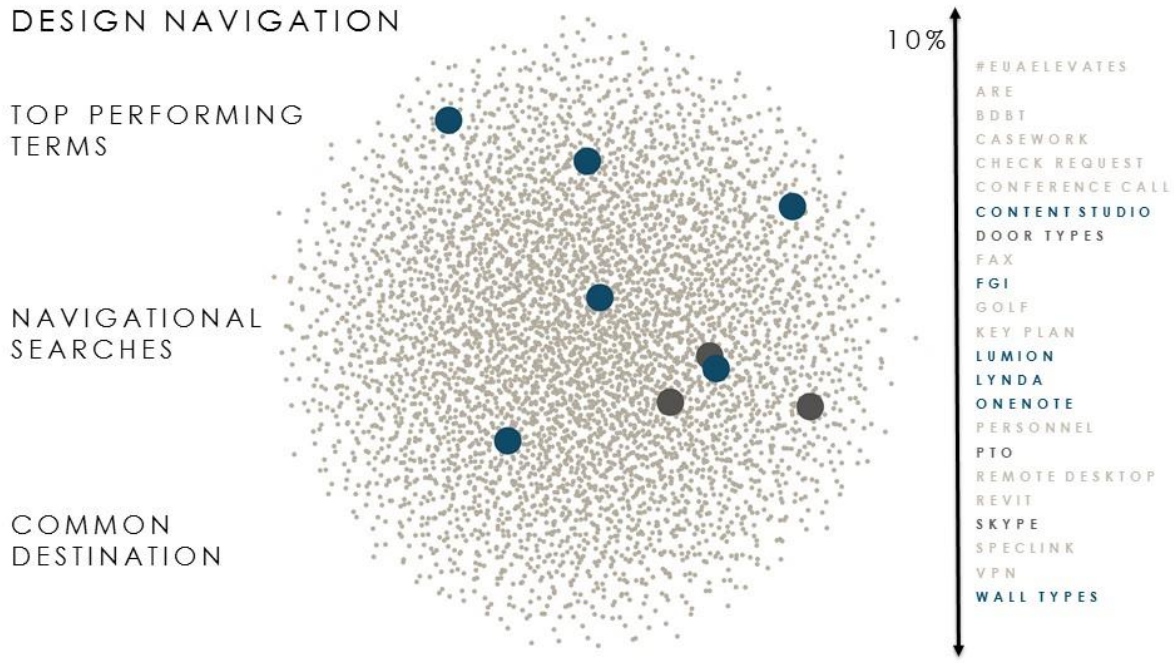

Figure 10: Top Performing terms exhibiting some navigational search behavior

The term 'FGI' [Facility Guidelines Institute] is an example of a navigational search in our dataset. The below example shows sessions from different users that last less than thirty seconds and all end with the same destination. This tells us that the community is using the search box to navigate to this particular wikipage as a short cut through the existing design.

Figure 11: Multiple session view for the term FGI [Facility Guidelines Institute]

In contrast, there are queries that exhibit some of the behaviour of a navigational search but do not fit all of the criteria. ‘Door Types’ in the example below is a term that exhibits very short search sessions, and some even arrive at a relevant resource, but there is no consistency across

all of the searches. It is the consistent access of a resource via short search sessions that denotes navigational search behaviour.

Figure 12: Multiple session view for the term 'door types'

Useful? Sure, a more efficient and relevant design is always desirable, still, we only have a two-legged chair - let's keep building.

4.1 CREATE CONTENT

Search data can be used to inform content creation. We can use this dataset to inform content creators about gaps in the knowledge base. This is where our strategy becomes more in-depth and we need to look past the top ten percent of search queries and begin an exploration of the entire dataset. We are looking for a validation of the resource cost to content creators and thought leaders in developing new material for the corporate knowledge base.

First, we identify underperforming terms or terms with a search score of zero, where the seeker abandoned the search due to lack of relevant results. Once a term is identified we use the entire dataset in a sensemaking task, finding semantic relationships between terms. Sensemaking consists of gathering the data, re-representing the data in a schema to aid analysis, and then the manipulation of this data to make those connections (Pirolli and Card 2005, p. 2). Using a visual representation once the data is conditioned to these ends can help us see the true distribution of the terms semantically related to the originally identified query. The distribution of this new semantic network should include terms with a number of instances. Non-relevant results for the information seeker on one side of the spectrum and unique queries, or queries that do not resemble any other and are located in the 'long tail' of the dataset on the other. Query complexity, normally a gauge of the information literacy of the seeker, can also be a marker found among unique queries that infer frustration by the user. A sophisticated searcher might, when faced with a sufficient amount of non-relevant results, begin to pull out everything in her information literacy toolkit, Boolean and other query operators for example (Jansen 2006, p. 419).

It is understood that the goal of information retrieval systems is to respond quickly to queries and present search results of maximum relevance to the seeker (Hawking 2004, p. 2). Enterprise information retrieval systems are no different but they often allow the knowledge manager to employ an advantage over search engines out in the wild. First, the search dataset is from a homogenous group with a more-or-less unified goal such as the firm's mission, vision, markets, and the providing of its services. Second, the knowledge manager will have context about these queries from her normal engagement with the community of seekers that an analyst of wild search data cannot possess. It should be stated here that any use of a search query dataset should be treated with a strong ethical baseline that preserves the privacy of the user, regardless of her status as an employee for the organization and who the data technically belongs to. A wild query log will contain data such as a user's IP address, a time stamp, the query, the user's OS or browser, etc. (Cooper 2008, p. 2). An enterprise dataset, depending on the content management system that is collecting it, will have more granular data points like a username, for instance, which can then be cross-referenced against an employee database, revealing a great deal more. Some of this might be useful, for instance, identifying subsets of the query data that are coming from a particular office or business unit. The further one drills down into the cross-referenced data, however, the closer one gets to a breach of basic rights to privacy. As we perform a deeper semantic analysis with the specific purpose of creating new content for our knowledge base, we walk close to that ethical border and should remain vigilant in the face of requests for more details or more granular connections.

Once a subset of semantically related terms has been identified and a gap in the knowledge base is verified, we can set about creating new content to fill the hole. Web 2.0 tools such as blogs, microblogs, wikis, and social newsfeeds are an excellent way to quickly close that gap (Kotval and Burns 2013, p. 78).

4.2 CREATE CONTENT - METHODOLOGY

The first step in this example consists of identifying an underperforming search term. This term may or may not be in the top tier of search queries so it is not necessary at this stage to cross reference a first choice against the distribution of query instances that we have produced with Tableau. One need not perform any further analysis, contingent on the specifics of the KM mission requiring the action. One could create new content to accommodate any underperforming search term without additional analysis. If, however, a knowledge manager requires a greater justification on the resource cost of content creation, which can be significant, then she should then look to the rest of the dataset for terms that are semantically related to her initial choice.

For the results below the identification of semantically related terms was done in a largely manual fashion using the 'Find' function on a copy of the dataset in Excel. Instances of terms that contained either 'ANSI' or 'ADA' were counted as semantically related. These relationships, and similar ones, can be found via more sophisticated or automated methods, but the scale of this exercise allowed for a manual count using a common tool.

4.3 CREATE CONTENT - RESULTS

As stated, we need to identify a gap in our knowledge base first and we do that by looking for underperforming terms. In this example the term is 'ANSI ADA.'

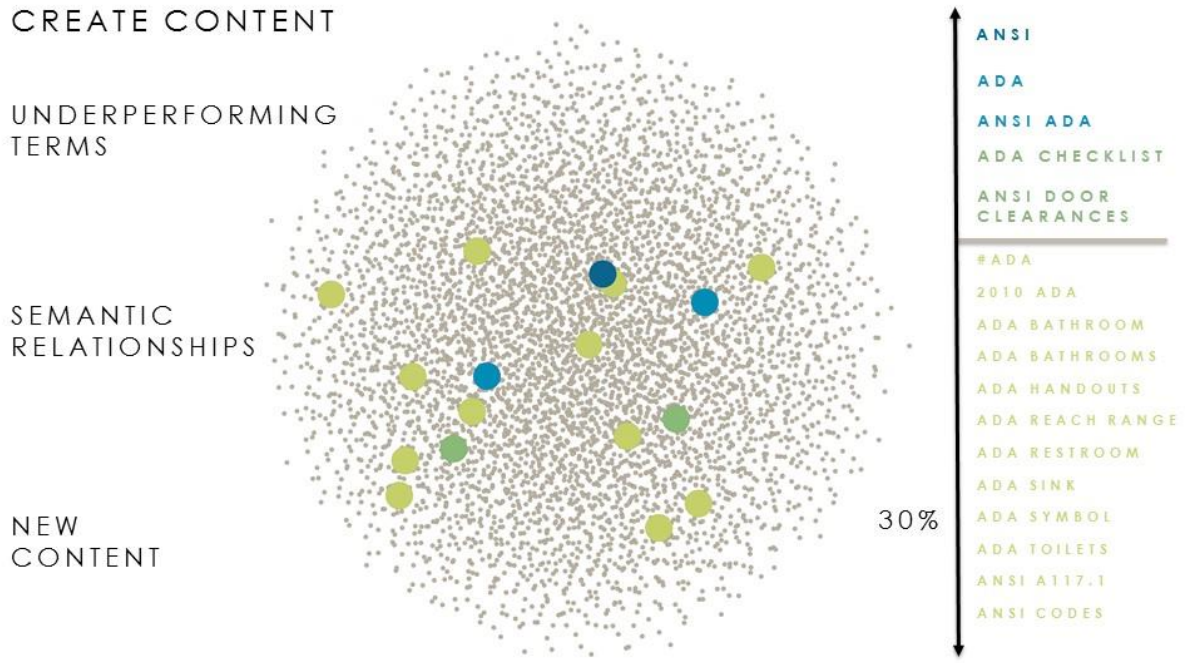

Figure 13: Terms semantically related to ANSI ADA across the top 30% of the dataset

The case for expending the resource cost in creating new content was strengthened through a manual search for terms semantically related to 'ANSI ADA'. Once this was completed it became clear that users were searching for content related to this term using a number of different queries. They were having difficulty finding relevant content. This was an action item, a gap that needed to be filled.

A post for the intranet newsfeed, an open question, on what the best source of information related to 'ANSI ADA' was the quickest route to filling the gap. In this example an answer came within minutes from a few community members familiar with the resources relevant to the query. The post now contained links to both external and internal documents and the next time the term was searched by a member of the firm, the gap in the knowledge base would be filled.

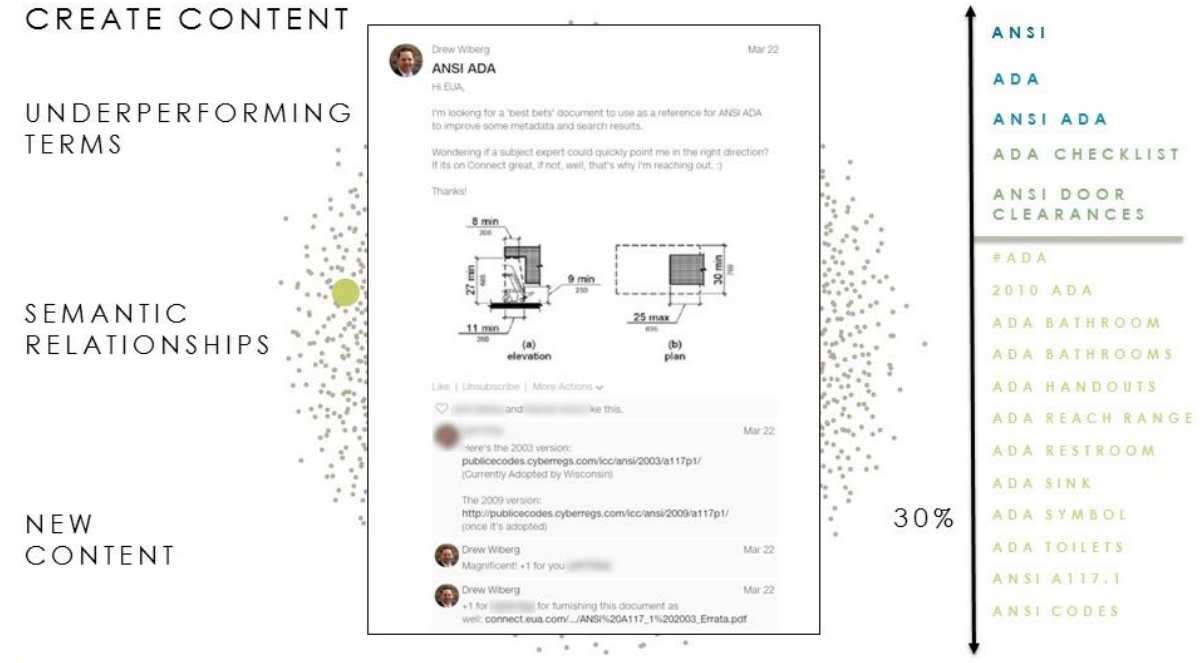

Figure 14: Newsfeed post requesting resources related to the term 'ANSI ADA'

Now, we are getting somewhere, right? We have a three-legged chair and the beginnings of a stable KM strategy. Let's look at one more use for our dataset to round it out.

5.1 DEVELOP TRAINING

As a final leg of our data-driven knowledge management strategy we can leverage our search query dataset to inform the direction and administration of a corporate university program. The purpose of the program is to provide problem-oriented learning that is applicable to architectural project work as well as more culturally directed, lifestyle-oriented, learning opportunities. Most training courses are developed as hour long blocks but it has been suggested that courses be developed as microlearning events of a half-hour or less. These courses would be targeted towards teaching / learning one technique, tool, or wrestling with a specific problem. The below method of informing curriculum could easily support the implementation of microlearning events. Research in microlearning has shown that this format is more digestible for a workforce that is increasingly populated with digital natives. Interactive problem solving that conveys information quickly, uses graphical approaches over text, encourages team activities, frequent rewards and the use of actionable tools are some of the approaches that are resonating with a younger workforce as former generations reach retirement age (Charles 2015, p. 55). Using search queries as reference points that reveal different perspectives for a user or corpus of user's information needs can reveal trends in subject topics and the multiple facets that a particular subject can be broken into and still found as relevant (Zhang 2007, p. 2). Coordinating the use of search query data with microlearning events recognizes the emerging postmodern fragmentation of information resources and their use in a quick-moving, project focused environment (Kruger 2012, p. 2). Search query data is, by nature, a reference of self-directed microlearning events. Reflecting this self-directed learning through the administration and offerings of a corporate education program reinforces the initial efforts put forth by the corporate community, deepens the body of tacit knowledge already obtained by community members, and recognizes the user as the expert on her own information needs as opposed to the institution's opinion on what she needs to learn.

The majority of learning today, in general, takes the form of informal microlearning using the internet as a knowledge base (Kruger 2012, p. 8). Modern educational platforms such as Udacity (http://www.udacity.com) offer their courses online and designed as easily digestible snippets. A corporate university has the added benefit of serving a group with largely the same business goals and a unified cultural context. By emulating the successful educational methodology of newcomers like Udacity, leveraging that shared context, and utilizing search data to improve and inform the direction of course offerings at no added resource cost to the user, the corporate university places itself in a position weighted towards success and engagement. It is important to mention the criticism against search query log analysis at this point, namely that this analysis technique does not get at the underlying information need of the seeker, whether or not they were satisfied with the results of their search, and that they only reveal a sequence of actions taken by the seeker when using the search engine (Jansen 2006, p. 410). These are valid critiques of this analytical method but they begin to fall apart when applied to an enterprise search environment, especially when the method is applied by a dedicated knowledge management professional within the organization. It is a knowledge manager's core duty to understand the community she serves and the microcommunities that make up the larger whole. Through participant observational techniques a knowledge manager gains an explicit understanding of a firm's culture and the information needs of the learners within. With this background she can look at search sessions with minimal information - sophistication of query, depth of results accessed, length of the session, etc. - and with no additional information on the actual user, can make connections using her understanding of the cultural context of her firm and the current project work. She can then use those intersections to administer a highly focused and effective educational program for her community.

While the entire dataset is relevant to this method, the type of query that is weighted towards this use is the exploratory search. An exploratory search is multifaceted, has open-ended search goals or vague queries using broad terms, and is obvious through the action of the seeker that it is a cognitively complex action with sometimes convoluted search paths (Athukorala et al 2015, p. 5). This behaviour suggests that the information seeker is not only looking for an information resource to fulfil her need, but is looking for multiple resources that might inform a single need from different perspectives. The knowledge manager viewing these types of sessions will see broad term queries followed by semantically related queries of either broader or narrow scope as the seeker navigates the knowledge base in her exploration. Employees at a firm with a well-developed intranet will typically fill the search query log with sessions seeking company policies, financial or project information, client histories, or administrative logistics (Hawking 2004, p. 2). Exploratory searches will stand out against this behaviour, exhibiting iterative actions and increasing complexity of session actions and query formulation (Association of College and Research Libraries [ACRL] 2016, p. 7). In the exploratory search process, information seekers become learners developing more and more sophisticated information literacy. They navigate between results, recognizing the different formats in which they are packaged, developing an understanding of how their choices impact the relevance of their results (ACRL 2016, p. 5). These searches, this process, are guideposts for the knowledge manager analysing the dataset, allowing the community to passively co-design the direction of their internal university's curriculum.

Additionally, research suggests that the type of result selected during a search session can inform the administrator of the educational program as to the nature of the instructional method best utilized for a specific subject. If, within the intranet's information resources, wikis are often selected, this suggests that active, collaborative and visual learning might be the best method. If the results are weighted towards blogs or newsfeed posts, this suggests a more

reflective, communicative, lecture learning style might be preferred by those self-instructing on that same subject. There is likely a correlation with the reduction of focused, directed queries and the increase of exploratory tasks as a knowledge base becomes more fragmented and a project worker's tasks become more specific (Kruger 2012, p. 1). A general increase in exploratory searches can indicate not only the direction an internal university should take, but the aggressiveness that its administrators promote and implement its learning events.

5.2 DEVELOP TRAINING - METHODOLOGY

In the process of obtaining the below results, two datasets were used. The first was our familiar search query log and the second, a list of historical course offerings for the past three years at EUA. These datasets were not immediately relevant towards one another, so a manual process of subject analysis was undertaken for both. Using the course offerings and tacit cultural context as a guide, the courses were grouped into larger subjects. In this example, those subjects were directly related to the practice areas recognized by the firm. As a side note, the intranet is also designed to correlate with the practice areas, making this structure that much more relevant to the administrators of the internal university program.

Following this exercise, which took an estimated four hours of the knowledge manager's time, the same structure was imposed on the query dataset. In the interest of thoroughness, every query was placed into a larger category. There was some replication of data as queries were sometimes vague and relevant to more than one subject area. This exercise, while it might seem daunting, was done in less than six hours of manual analysis. Again, this process could be automated if it was decided that the scale or frequency of the exercise was to be increased. No attempt was made to analyse the differences between individual sessions (exploratory vs. look-up, for example) or the differences between the results selected as relevant (wikis vs. newsfeed posts) due to the time constraints of the exercise.

Once both datasets were analysed and the data contained in subject areas now related to one another easily, the percentage of data under each subject was then calculated allowing the two datasets to be compared.

5.3 DEVELOP TRAINING - RESULTS

Using the above methodology, user needs in a given subject were identified. This allowed for educated suggestions on how to fill the gaps in the internal university's curriculum.

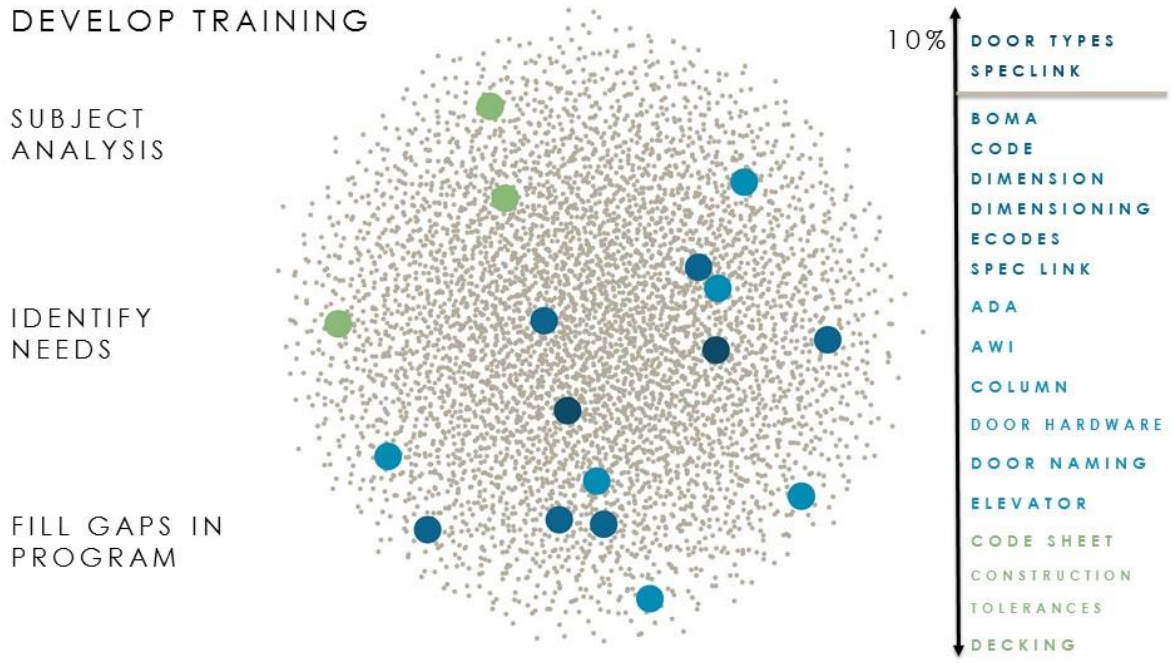

Figure 15: View of search terms that fall within the subject of 'Implementation'

Three years of course offerings were scrutinized and boiled into larger categories, the same being done with the dataset of search terms. The resource cost of this exercise had been previously justified by a request from leadership to improve the internal university program.

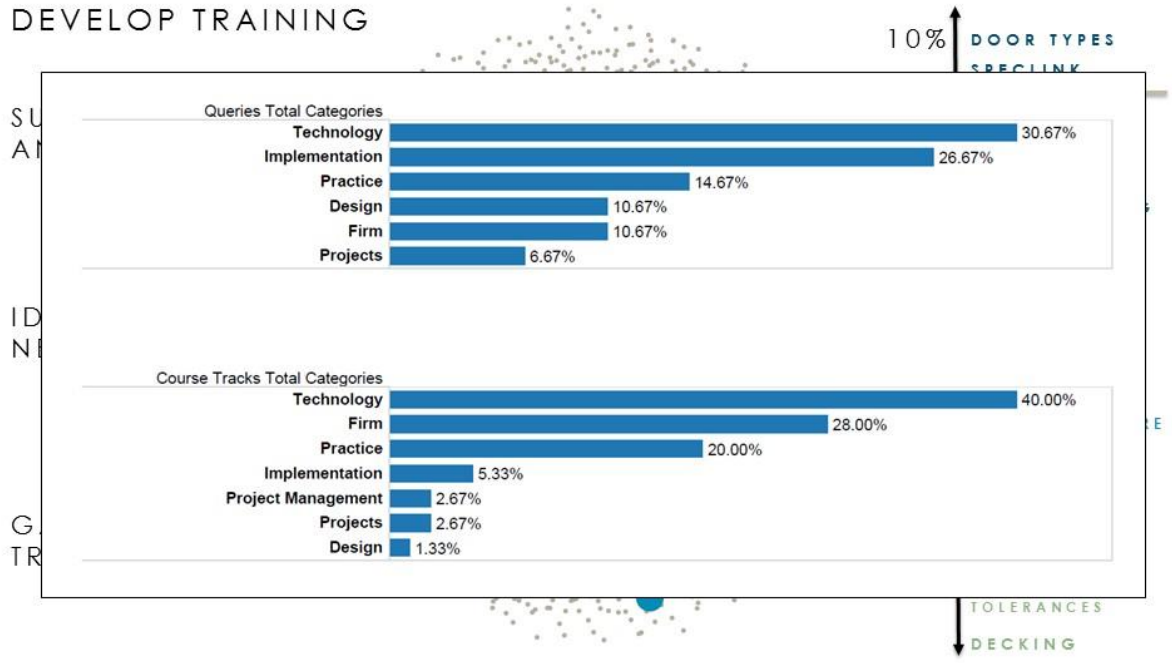

Figure 16: Tableau visualization of the comparison of selected subjects

An example of a gap found after analysis was in the category of Implementation. 26.6% of the search terms in the search query dataset somehow related to the subject of Implementation, yet only 5.33% of the historic course offerings fell into that same category.

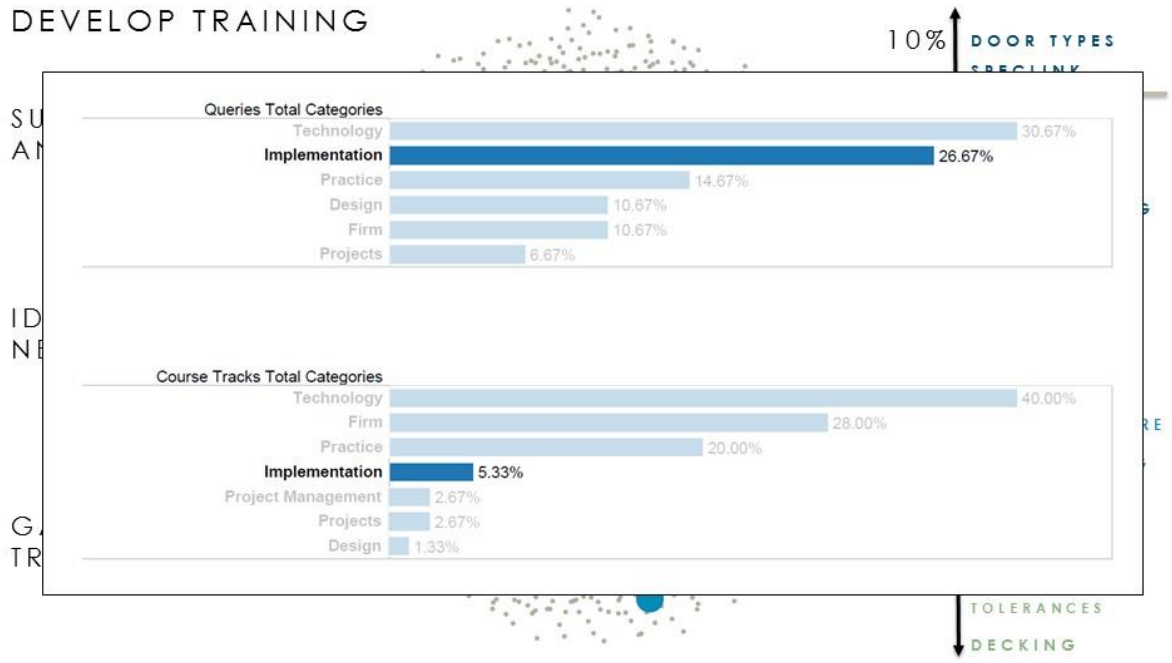

Figure 17: Highlight of difference between percentages in the Implementation subject

After the gap in the education program was verified via the data analysis, leadership requested a list of terms used by information seekers that related to the subject of Implementation. This step had already been completed early in the analysis, effectively auto-populating a list of categories related to Implementation that could be taught through the training program.

Following this last step of analysis, we now have a solid KM strategy that can help connect people with each other and with the information resources they need. A chair with four stable legs that allows us to sit at the same table as our community and leverage their conversations with the search engine and with each other to connect them with the knowledge they need in an efficient way.

6.1 SUMMARY

In summary, search data logs can be a very effective tool for a knowledge manager looking for ways in which to impact her firm with little to no resource cost to the project workers there. Rather than being a static record of queries, search data can, in the hands of a knowledge manager with sufficient embeddedness and an understanding of the firm's culture, be a window into the connections that need to be made. These are connections between people and the knowledge base (improving search and navigational design) or connections between different people at the firm (creating content and developing training). The knowledge manager can achieve a deep view into the information seeking behaviour of her community and guide her work to be more relevant, more efficient, and best of all, nearly invisible to those it impacts.

Bibliography

1) APCQ (2013) Transferring and applying critical knowledge. pp 1-221.

2) Association of College and Research Libraries (2016) Framework for information literacy for higher education. pp 1-18

3) Battleson B, Booth A and Weintrop J (2001) Usability testing of an academic library website: A case study. The Journal of Academic Librarianship (27, 2) pp. 188-198.

4) Blanke T (2005) Ethical subjectification and search engines: Ethics reconsidered. International Review of Information Ethics (3) pp. 34-38

5) Brantley S, Armstrong A and Lewis K M (2006) Usability testing of a customizable library portal. College and Research Libraries. pp 146 - 163

6) Charles L H (2015) Using an information literacy curriculum map as a means of communication and accountability for stakeholders in higher education. Journal of Information Literacy. (9, 1) pp. 47-61.

7) Cooper A (2008) A survey of query log privacy-enhancing techniques from a policy perspective. ACM Transactions on the Web (2, 4) pp. 19:1-19:27.

8) Google, Inc. (2013) Google search appliance. Ch. 3. Customizing the user interface. pp 94-119

9) Han H, Jeong W and Wolfram D (2014) Log analysis of academic digital library: User query patters. iConference 2014 Proceedings. pp. 1002-1008.

10) Hu, Y., & Shi, L. (March 01, 2015). Visualizing large graphs. Wiley Interdisciplinary Reviews: Computational Statistics, 7, 2, 115-136.

11) Jarrett, C., Quesenbery, W., Roddis, I., Allen, S., & Stirling, V. (January 01, 2009). Using Measurements from Usability Testing, Search Log Analysis and Web Traffic Analysis to Inform Development of a Complex Web Site Used for Complex Tasks. Lecture Notes in Computer Science, 5619, 729-738.

12) Kruger N (2012) Micro-E-Learning in information literacy. IFLA 2012. pp. 1-10.

13) Lieberman M (2014) Visualizing big data: Social network analysis. Multivariate Solutions. pp. 1-23.

14) Liu S (2008) Engaging users: The future of academic library web sites. College and Research Libraries. pp 5-27.

15) Manzari L and Trinidad-Christensen J (2006) User-centred design of a web site for library and information science students: Heuristic evaluation and usability testing. Information Technology and Libraries. pp. 163-169.

16) Maxwell G and Beattie D R (2004) The ethics of in-company research: An exploratory study. Journal of Business Ethics. (52, 3) pp. 243-256.

17) Murray G C and Teevan J (2007) Query log analysis: Social and technological challenges. ACM SIGIR Forum (41, 2) pp 112-120.

18) Plaisant C (2004) The challenge of information visualization evaluation. IEEE Proceedings of AVI 2004. pp 1-8.

19) Quesenbery W, Jarrett C, Rodding I, Stirling V, Allen S (2008) Search is now normal behaviour. What do we do about that? Usability Professionals' Association. pp 1-12
Richardson M (2008) Learning about the world through long-term query logs. ACN Transactions on the Web (2, 4) 21:1-21:26.

20) Zhang J (2007) Visualization for information retrieval. pp 1-5.