The best of both worlds: Highlighting the synergies of combining manual and automatic knowledge organization methods to improve information search and discovery in oil and gas enterprises.

Paul H. Cleverley
Department of Information Management, Robert Gordon University, Garthdee Road, Aberdeen AB10 7QB Email: p.h.cleverley@rgu.ac.uk

Simon Burnett
Department of Information Management, Robert Gordon University, Garthdee Road, Aberdeen AB10 7QB Email: s.burnett@rgu.ac.uk

Abstract

Research suggests organizations across all sectors waste a significant amount of time looking for information and often fail to leverage the information they have. In response, many organizations have deployed some form of enterprise search to improve the 'findability' of information. Debates persist as to whether thesauri and manual indexing or automated machine learning techniques should be used to enhance discovery of information. In addition, the extent to which a Knowledge Organization System (KOS) enhances discoveries or indeed blinds us to new ones remains a moot point. The oil and gas industry is used as a case study using a representative organization. Drawing on prior research, a theoretical model is presented which aims to overcome the shortcomings of each approach. This synergistic model could help to re-conceptualize the 'manual' versus 'automatic' debate in many enterprises, accommodating a broader range of information needs. This may enable enterprises to develop more effective information and knowledge management strategies and ease the tension between what are often perceived as mutually exclusive competing approaches. Certain aspects of the theoretical model may be transferable to other industries, which is an area for further research.

Introduction

Oil and gas exploration seeks to identify and model hydrocarbon resources through geoscientific methods. Exploration wells can cost over $100Million in deep water (Blackman 2012) and typically have a 30% chance of success (Oil and Gas UK 2011). It is therefore critical that all relevant information is included.

A review of surveys across all business sectors indicates 24% of a business professional's time is spent looking for information (Chui *et al.* 2012, Doane 2010, Outsell 2005, Feldman *et al.* 2005, Lowe *et al.* 2004, Adkins 2003, Delphi 2002, Feldman and Sherman 2001). Much lower figures (9-14%) have been reported from observational studies in organizations (Robinson 2010, Majid *et al.* 2000) and much higher figures (40%) reported in the oil and gas industry (Hills 2014, Chum *et al.* 2011). A review of surveys indicates that 48% of organizations felt search was unsatisfactory (Norling & Boye 2013, Mindmeter 2011, Doane 2010, Microsoft & Accenture 2010, Feldman 2009, AIIM 2008, Feldman *et al.* 2005, Tonstad & Bjorge 2003).

Executives indicate missed opportunities caused by failing to leverage information effectively in the oil and gas enterprise could represent as much as 22% of annual revenue (Oracle 2012). Acknowledging this significant opportunity cost, Rasmus (2013) proposes the *Serendipity Economy*, where discovery of information can produce major leaps in value that cannot be predicted. Exploiting and using information to make better decisions and improve performance are the goals for Knowledge Management (KM).

Causal factors for enterprise search performance are numerous, including information silos, search expertise, governance and technology issues (White 2012, DeLone and McLean 2002). Data from search logs (Dale 2013, Romero 2013) and from practitioners (Andersen 2012, White 2012), indicate issues exist with enterprise search. One issue is the *vocabulary problem* where two people will not choose the same name for the same concept 80% of the time (Furnas *et al.* 1987), causing a mismatch between the search terms used and the information sought. This leads to challenges for enterprise search in finding precise information and recalling all relevant information. Another issue is the minimal use of faceted search and categories which rarely stimulate serendipitous encounters (Cleverley and Burnett 2015a). Despite major investments, dissatisfaction with enterprise search is widespread (White 2014, Norling and Boye 2013).

The role and concomitant benefits of thesauri and manual indexing as well as automated machine learning techniques in information discovery is a source of ongoing debate. While this topic is well developed within the literature, it is far from being addressed conclusively. Collins and Porras (1997, pg. 10) describe the decision making process of visionary companies in terms of *"the tyranny of the OR*, *the genius of the AND"* when coping with contradictory forces. Is this a philosophy to apply to enterprise information (knowledge organization) with respect to manual and automated methods?

Furthermore, Knowledge Organization Systems (KOS) themselves may act to reveal, or conversely obscure information discoveries. Given these issues, there is a need to assess how manual and automated Knowledge Organization (KO) techniques might support information search and discovery. This research therefore reconsiders these issues within the context of the oil and gas industry, with the explicit intention of developing a synergistic model which encompasses the main benefits of each approach into a 'best of both worlds' scenario. The following research questions were identified to fulfill the aim of the research, the rationale for their inclusion is presented in the literature review:

Q1. To what extent can a thesaurus be enhanced through automated techniques?

Q2. What is the value of auto-categorizing content that is already manually classified?

Q3. To what extent can manual and automated KOS techniques be combined in a search user interface to stimulate serendipity?

The next section reviews the literature with a focus on oil and gas, followed by the methodology. The results are presented with discussion to help the reader better understand the findings and limitations. The paper concludes with the presentation of a theoretical model, areas for further research and implications for theory and practice.

Literature Review

This section presents a critical review of the academic and practitioners literatures relevant to the research, guided by a conceptual model (Figure 1).

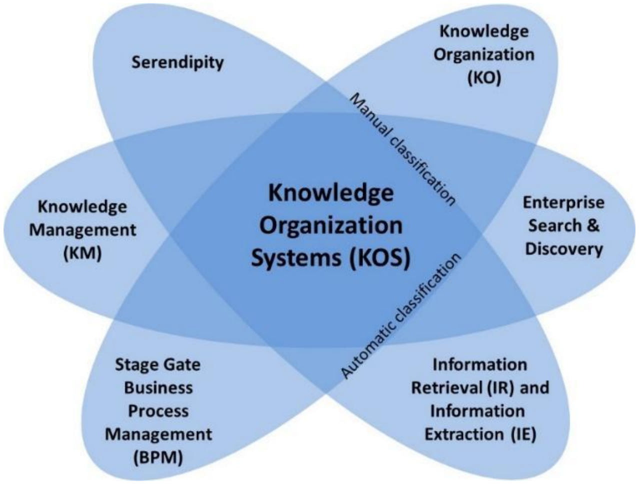

Serendipity
Knowledge Organization (KO)
Knowledge Management (KM)
Knowledge Organization Systems (KOS)
Manual classification
Enterprise Search & Discovery
Automatic classification
Stage Gate Business Process Management (BPM)
Information Retrieval (IR) and Information Extraction (IE)

Figure 1 – Conceptual model to guide the reader through the literature review.

The literature review provides a background to the areas under research from both academic and practitioner standpoints, identifies how the literature has informed the research questions, presents gaps in the existing literature and emphasizes how this research addresses those gaps, and highlights areas of input into the final theoretical model.

Knowledge Organization (KO)

Knowledge Organization (KO) expresses and imposes a particular structure of knowledge (a 'view of reality') behind collections of information (Ohly 2012). This reality is socially constructed: what is reality for one group may not be for another (Berger and Luckman 1966). Hjorland (2008, pg. 86) offers a holistic definition of KO, encompassing the broader *social* division of mental labour, to the narrower *intellectual* activities, *"..such as document description, indexing and classification performed in libraries, databases, archives etc. These activities are done by librarians, archivists, subject specialists as well as by computer algorithms"*. Hjorland continues, "*Library and Information Science (LIS) is the central discipline of KO in this narrow sense (although seriously challenged by, among other fields, computer science)"*. This alludes to the tension that exists between Library and Computer Science.

Recent evidence from organizations (Quaadgras and Beath 2011) contradicts the definition made by Hjorland that KO is the preserve of information specialists. Corporate library or information center functions have traditionally focused on the centralized manual indexing of information using KOS, with indexes under their stewardship (Heye 2003). The growth in digital information creation has led to the breakup of these gatekeeping services and the centralized manual indexing model. Zeeman *et al.* (2011) found government libraries plan to deploy, *"high-end thesaurus and ontology tools.. to work with structured and unstructured data for decision-making research".* This provides evidence of how some corporate librarian skills and services are changing.

Classification and categorization can be achieved manually (by creator or mediator) or automatically through supervised/semi-supervised machine learning. The use of the terms *classification* and *categorization* have been (and continue to be) used interchangeably by practitioners and can cause conceptual misunderstandings. Simplifying, classification organizes information to mutually exclusive non overlapping classes, whilst categorization is more flexible, recognizing similarities across entities enabling information to be organized into one or more categories (Jacob 2004). Applying this to a 'typical' oil and gas document, classification may involve assigning an item to a single Document Type it *is* a 'Well Proposal'. Whilst categorization may include assigning the document to be *about* oil and gas well '33/4b-5' and 'light tight oil'. Classification and categorization typically need an existing set of classes/categories like a taxonomy or authority list, whilst 'tagging' is also used to refer to the process of adding terms which may include those from outside controlled vocabularies to emphasize prominence (Hedden 2013).

Knowledge Organization Systems (KOS)

Hodge (2000, pg.1) defines Knowledge Organization Systems (KOS) as including, *"classification and categorization schemes that organize materials at a general level, subject headings… and authority files that control variant versions of key information such as geographic names and personal names. Knowledge organization systems also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies."* This definition is adopted for the research study, including automatically generated associative thesauri that involve no manual input.

Zeng (2008) arranges KOS types in order of increasing sophistication, by both structure and use cases (eliminating ambiguity, controlling synonyms, establishing relationships and presenting properties). A corporate taxonomy language that fits the oil and gas organization is seen as critical to ensuring content governance, navigation and retrieval. This is evidenced by multinationals such as RepsolYPF (Salmador-Sanchez and Angeles-Palacios 2008) and Statoil (Munkvold *et al.* 2006), small independents such as Southwestern Energy (Caballero and Nuernberg 2014) and Apache Energy (Rose 2010), National Oil Companies such as Petronas (Noor and Yassin 2006), service companies such as Baker Hughes (Hubert 2012) and Governments such as the Ministry of Oil and Gas in Oman (Alyahyaee 2012).

From an information search engine perspective, KOS can mitigate the *vocabulary problem* when converted into machine readable forms through Knowledge Engineering (KE) techniques (Preece *et al.* 2001). In contrast, oil companies such as RepsolYPF found commercial thesauri needed constant maintenance and had poor coverage, they used Computer Aided Knowledge Engineering (CAKE) methodologies to generate a thesaurus (Salmador-Sanchez and Angeles-Palacios 2008) but no methodological detail was provided. Concept hierarchies can be created automatically through text clustering (Palmer *et al.* 2001). Automated thesaurus creation and enrichment techniques from text corpora are well documented (Grefenstette 1994), although Velardi *et al.* (2012) stated it is virtually impossible to recreate complex domain specific taxonomies automatically from document content alone. This raises the question to what extent can oil and gas thesauri be enhanced using automated techniques?

As stated previously, there is a debate whether to use manual or automated methods to classify information. In a provocative debate held in 2015 by the UK chapter of the International Society for

Knowledge Organization (ISKO) the following question was posed *“This house believes that the traditional thesaurus has no place in modern information retrieval”*. It was argued that thesauri are no longer of value as searchers just want to type search terms in a search box. The media concluded, “*The pro search, anti-thesaurus motion was defeated resoundingly”* (McNaughton 2015, pg.3). Two separate questions may be conflated in the debate. Firstly, the extent to which people need to browse or navigate to information, compared to typing terms in a search box and secondly the extent to which thesauri are advantageous to help classify, categorize, find or discover information (through manual or automated methods).

There is also a debate on the extent to which manually created KOS provide benefits. According to Greenberg (2011 pg. 12), *“when knowledge structures are absent, the information system is generally considered sub-standard. KOS are a necessity: they inform and promote discovery, use and re-use of information”*. Greenberg contrasts this with, *“Benefits aside, we must also acknowledge that schemes may reinforce erroneous views, false perceptions and limit new discoveries.”*

A thesaurus provides a controlled vocabulary of terms that contain hierarchical and associative (Related Term (RT)) relationships. The hierarchical relationships, *‘is a’ Taxonomy* or *‘part of’ Meronymy* (Salthe 2012), are the backbone of a thesaurus. It is relatively straightforward to use these parts of the thesaurus for machine readable classification or categorization, although confusion has existed on the difference between ‘concepts’ (associative and hierarchical relationships are between concepts) and ‘terms’ (equivalence relationships are between terms) (Dextre Clarke and Zeng 2012). The automated use of RT associations is more problematic due to the inconsistent application and ambiguous nature of the relationships in many thesauri (Spiteri 2004, Tudhope *et al.* 2001). As emphasized by Stock (2010, pg.1951), “*Unspecific associative relations are of little help to our focused applications and should be replaced by generalized and domain-specific relations”*. Shiri *et al.* (2002) found evidence of thesauri being marginal and a substantial source of terms to take advantage of when searching.

Information Retrieval (IR)

Manning *et al.* (2009, pg.1) describes Information Retrieval (IR) as, *“..finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”* Marchionini (2006) differentiated between two different goals of searching which need to be addressed by IR. Firstly ‘lookup (known item)’ search, where there typically is a right answer or search result. Secondly, ‘exploratory’ search to learn/investigate, where the outcome is uncertain, multi-faceted and delivers many results. Morville and Rosenfeld (2006) identified additional seeking models of ‘exhaustive’ (a form of exploratory search) and ‘re-finding’ which applies to both.

The discipline of Text Analytics (TA) uses Natural Language Processing (NLP) techniques to enhance IR (Grimes 2014). These NLP techniques attempt to understand text in the same way as humans do, catering for the creativity (and ambiguities) in grammar (Manning and Schutze 1999). In combination they are often referred to in an IT context as ‘search and discovery’.

A number of IR approaches to contextual search are put forward by Bhogal *et al.* (2007), including personalization, language models, ubiquitous computing, user background and task based context.

Language models include the area of Automated Query Expansion (AQE) which expands the original users query with more words that may better represent the original intent and can be divided into corpus dependent and corpus independent techniques. This can be achieved linguistically or statistically through creation of an 'associative thesauri' (Carpineto and Romano 2012, Luke *et al.* 2012). Lykke and Eslau (2010) reported recall improvements of over 100% in a pharmaceutical enterprise through thesaurus enhanced natural language (full text) search compared to keyword based natural language (full text) search. Other corpus independent techniques include the statistical analysis of Wikipedia (Peng *et al.* 2009) to linguistic based KOS such as Wordnet (Miller *et al.* 1990) to understand synonymous meanings of common English words. One drawback of Wordnet is its lack of technical domain specific terminology (Navigli and Velardi 2002). Examples of corpus dependent techniques include the use of term co-occurrences for weighted query expansion and techniques such as latent semantic analysis (Mikolov *et al.* 2013, Landauer and Dumais 1997). These types of search are often termed *semantic search*.

In a comparison of empirical results, Carpineto and Romano (2012, pg. 36), state, *"linguistic techniques are considered less effective than those based on statistical analysis..., but statistical analysis may not always be applied (e.g., when good expansion terms do not frequently co-occur with the query terms). Of the statistical techniques, local analysis seems to perform better than corpus analysis because the extracted features are query specific"*. Carpineto and Romano (2012, pg. 41) conclude, *"Hybrid [linguistic and statistical] methods achieved the best results on the experimental benchmarks and seem, in principle, more robust with respect to variation of queries, document collections and users"*. Combining manually generated KOS with statistical techniques could mitigate the shortcomings of both individual methods.

Thesauri (Shiri *et al.* 2002) and ontologies (Prince and Roche 2009) have been used extensively for AQE. In this context, Ontology is a conceptualization which aims to represent typically a single view of reality. Solskinnsbakk and Gulla (2008) found issues relating to the names used for concepts in oil industry ontologies (ISO 15926) compared to the 'everyday parlance' language used in documents, making it problematical to use ontologies directly for AQE. To overcome these issues, they merged an oil and gas glossary (Schlumberger 2008) with the ontology to improve search recall through statistical AQE. Search recall was improved but at the cost of precision, although only 7 queries and 130 Internet documents were used in the study. Conversely, Cleverley (2012) applied a large oil and gas taxonomy to auto-categorize a corporate library containing 170,000 records that had already been manually classified to the 'whole'. The study used 50 subject based search queries and concluded that search recall could be improved by 43% (addressing the vocabulary problem) without a major loss in average precision. The study was limited in that it did not include statistical techniques.

Enterprise Search

Information searching behaviour is a subset of information seeking behaviour (Wilson 2000). A container of information includes any physical artefact, book, box, CD, record (data or document form) or person. Classification and/or categorization of this information can produce metadata which can be searched and used to boost search ranking, in addition to the textual content held within the containers (if electronic).

Enterprise information can be of Temporary Value (IoTV) or long term value (Record). The KO techniques used to organize these two contexts may be very different, based on the accuracy required and audience (IoTV tends to be known to its current audience (Khoo *et al.* 2007)). The future audience for a record is not always known, hence the need for more context (metadata) to aid future discovery and provenance. For these reasons, shared folder taxonomies/folksonomies are often used as the basis to organize IOTV, whereas additional metadata is often required for records.

Enterprise search typically refers to IR technology which automatically indexes enterprise content (including web pages, documents and people expertise profiles) providing a single place for staff to search without necessarily knowing where content is located (White 2012). Some oil and gas search deployments index both structured and unstructured information integrating through meta-models and vocabularies and also enable spatial (map based) search (Demartini 2009, Behounek and Casey 2007, Palkowsky 2005).

In enterprise search most queries are single word (lookup) and often portrayed as not working well compared to Internet search engines (Andersen 2012). The crowd using search in enterprises is very small compared to the Internet, hampering the effectiveness of using statistical crowdsourced usage data for all but the most common of search queries. Many users want enterprise search engines to work like Internet search engines, but may also be oblivious to the relevant content that can be missed during exploratory search tasks even using Internet search engines (Skoglund and Runeson 2009).

Lookup (known item) search (Marchionini 2006) is likely to account for between 80-90% by volume of all enterprise search tasks (Stenmark 2006) including accurately locating definitive documents (e.g. *End of Well Report for oil well 110/4b-5*). Exploratory search is open ended, (e.g. *what information do we have on this Bolivia license? What do we know about vuggy porosity in Dolomites*?). Different KO and KOS methods may be required to meet the browsing and searching needs for these two types of search goals, yet the KOS literature rarely differentiates between these two search goals. Connecting the work of Marchionini to the KO literature in a theoretical model may further understanding.

Geoscientists or engineers may not add many tags to their documents, evidenced by ExxonMobil *"Any capture of metadata that took more than ten seconds to saving a file was considered problematic"* (Garbarini *et al.* 2008, pg. 4). Difficulties may also arise discovering certain information, if browsing folder names is the only option available, evidenced by Shell (Lennon *et al.* 2012). This raises the question, what is the perceived value of auto-categorizing content that has already been manually classified in some way?

Serendipity

Innovation or creativity sparked by an unexpected seemingly random event is often called serendipity. Within organizations, the discovery of innovations and business opportunities is often serendipitous (Ghiselin 2010, Friedman 2010). Within the context of this research, serendipity is defined as the phenomenon of fortuitous unexpected information discovery and may be the consequence of immersion in information rich environments (McCay-Peet & Toms 2011) making unforeseen connections. Browsing can support creativity, whether the intent is purposive or exploratory in nature (Bawden 1986). A prerequisite to serendipity is a prepared mind (Foster & Ford 2003). Serendipity as a phenomenon is unlikely to be controllable; however, developing a capability that may lead to more opportunities for serendipitous encounters during information search is considered plausible.

Faceted Search and Information Extraction (IE)

Some search user interfaces incorporate faceted search to aid the browsing process (Hearst and Stoica 2009) improving the chances for 'serendipitous' discovery. Faceted search is an Interactive Information Retrieval (IIR) technique, displaying an overview of search results, inviting further interaction to filter information and can improve task performance (Fagan 2010). Low usage of faceted search (5-12%) has also been reported (Ballard and Blaine 2011, Niu and Hemminger 2010).

The deficiencies of current search User Interfaces (UI) to facilitate exploratory search, *"Current search engines do not sufficiently support exploration and discovery, as they do not provide an overview of a topic or assist the user by finding related information",* (Krestel *et al.* 2011, pg.393) and stimulate learning, *"[need for] higher levels of learning through the provision of more sophisticated, integrative and diverse search environments that support greater information immersion and more nuanced types of learning",* (Allan *et al*. 2012, pg.8) provide opportunities for KOS enhanced search UI research.

People can be attracted by visually salient colouring in user interfaces to highlight patterns which may otherwise remain obscured (O'Donnell 2011). Categories have been grouped by colour in faceted search (Hearst and Stoica 2009) and infographics (McCandless 2012). Where deployed, facet values are typically ordered by the most statistically frequent or most popular (Kaizer and Hodge 2005). Categories or tags displayed in faceted search that are representative of an information item (container) as whole, rarely contain intriguing or non-obvious associations (Cleverley and Burnett 2015b).

Information Extraction (IE) using search term word co-occurrence is one technique to introduce local context into search refiners '*what resources are nearby*' (Goker and Davies 2009, pg. 132). This technique uses IE to deconstruct sentences containing terms co-occurring around search terms in document text into their 'atomic concepts' (Smiraglia and van den Heuvel 2011) for use as search refiners. Crucially, this allows the same information container to be represented by *different* filter terms, depending on the search term used. Word co-occurrence filters may (Gwizdka 2009, Olsen 2007) or may not (Low 2011) aid discovery. Research studies indicate when browsing, the most intriguing or interesting concept or term associations may be the contextually unusual (Chuang *et al*. 2012) not the most statistically frequent. This raises the question, to what extent can manual and automated KOS techniques be combined in a search UI to stimulate serendipity?

Manual information classification

Digital information growth and stricter regulatory requirements has led to more federation of document publishing using Electronic Document Management Systems (EDMS) as part of Information Management (IM) strategies in the oil and gas industry (Gimmal 2013) evidenced by Marathon Oil (Smith 2012) and Chevron (Quaadgras and Beath 2011).

Manual based organizational IM can have several challenges. Firstly, it is unrealistic for all digital content to be manually assessed all of the time; fully optimized manual efforts are likely to be too expensive. Secondly, when end users are asked to classify and tag records, they may simply not do it, especially if it takes time (Garbarini *et al.* 2008). Finally, it is prone to the *vocabulary problem*, people will not always classify or categorize consistently, averages range from 91% (Faith 2011) to 46% (Magnuson 2014).

In the oil and gas industry, stage-gate processes typically help govern opportunities, prospects and projects as they pass through repeated execution, assurance and decision gates. It is crucial to be able to quickly locate the final version of a key document type (known item) that was created by (or used for) a particular stage gate process (Walkup and Ligon 2006). This stage gate structure lends itself to simple *pre-attributed* (pre-populated) process steps (folders), of deliverables required, underpinned by corporate taxonomies and authority lists. Documents added to these areas inherit this metadata enabling any document publisher (regardless of expertise or location) to 'drag and drop' documents to appropriate steps ensuring consistent application of metadata evidenced by Shell (Abel and Cleverley 2007). This may mitigate aspects of the *vocabulary problem*. This metadata can be used to both improve search result ranking and for graphical colour coded matrix dashboards supporting Business Process Management (BPM) for tracking and identification of missing deliverables for *proactive* information asset management.

Pre-attributed metadata inheritance methods have limitations. They use a small amount of controlled metadata to classify the 'whole' information item so predominantly support known item (lookup) search.

Automatic classification/categorization

Classification and categorization can be achieved through machine learning using linguistic rules, labeled training sets or a combination (Villena-Roman *et al.* 2011). In a study of legal document categorization, Roitblat *et al.* (2009, pg. 70) found machine categorization no less accurate than a team of reviewers, leading to the conclusion that *"machine categorization can be a reasonable substitute for human review*". Unsupervised machine learning organizes unlabeled information by latent structure and includes Topic Modeling (Yu *et al.* 2014, Meza 2014). Sidahmed *et al.* (2015, pg. 10) applied Topic Modeling on unstructured text in daily oil and gas drilling reports at BP to identify trending issues before they became serious, that were not apparent to engineers. Accuracy was raised as an issue, *"Lack of a drilling discipline concept dictionary..domain knowledge carries more weight during this part of the process"*.

Some studies place auto-classification accuracy at the 90% range (Sasaki 2008, Jacobs and Rau 1990) with practitioner heuristics indicating 70% accuracy (Faith 2011). Sasaki's study on Reuter's newswires used 9,603 training documents and only 11 target categories. Jurka *et al.* (2013) found accuracy rates of 65% using 4,000 training documents from the US Library of Congress. Accuracy rates of 60-90% were reported by the US Army, using 11,915 emails as a training set to auto-classify email to 54 records categories (Magnuson 2014). Depending on the content, category sophistication, subject matter expertise for rule creation and/or training data available (including its cleanness), it may be concluded that accuracy for machine learning auto-categorization typically varies between 60-90% (Miller 2014) and has 100% consistency. Practitioner based heuristics indicate 50-100 labelled training documents are typically required to give good results per category (Hedden 2013, Faith 2011).

Document type classes that require '*hard classification*' (binary classification to a single class) are probably the most challenging of machine learning tasks, accuracy percentages can be as low as 31%

(Painter *et al.* 2014). It can be difficult for automated approaches to work effectively without the necessary textual clues.

ConocoPhillips auto-categorized discussions and best practices which had already been manually organized by subject headings, allowing a depth of categorization/tagging that was unlikely to be achievable through manual methods. This was achieved through a manually created linguistic KOS aided by CAKE methods (Wessely 2011). Topic modeling has also been applied to enterprise lessons learnt systems to reveal hidden connections (Meza 2014). This contrasts with a lessons learned system deployed by ENI which focused on a single (manual) categorization approach (Piantanida *et al.* 2015).

In summary, for classification and categorization of large volumes of diverse information (e.g. discussions, news, emails) automatic methods are probably well suited. Where high levels of accuracy are required for key business deliverables and knowledge capture, manual methods are likely to offer better accuracy, particularly if pre-attribution of metadata can be used to enhance consistency where possible. Furthermore, combinations of manual and automated methods may offer additional business benefits. This review of the literature has led to the development of a number of research questions, in order to better understand how manual and automated KO and KOS approaches can be combined in a synergistic way to derive business value in oil and gas. The final research objective is to develop a theoretical model to explain the different KO/KOS approaches and how they support different search goals.

Study Methodology

A pragmatic research lens was chosen for this study. Ontologically, a pragmatist's view is that there is no objective reality: the 'truth' is that which works. It *"provides a framework of intellectual resources and rules for navigating our way in the experiential world in which we are embedded"* (Martela 2015, pg.17). Epistemologically, pragmatic research leads to warranted assertions through the process of inquiry linked to practical relevance – a need to act. It is a way of clarifying ideas by following the practical consequences of those ideas, ultimately favouring one idea over another. Pragmatism can view different theories as complementary some may work better than others for certain purposes and contexts. A mixed methods research design was used to collect and analyze both qualitative and quantitative data, with an approach based on grounded theory (Strauss and Corbin 1998) to thematically map nuances and comments.

Organizational data sampling and collection

The oil and gas industry is the case study, an exploration department in a large oil and gas company was the unit of analysis for the research. The organization was chosen because of its size as some surveys identify search as being more difficult in larger companies (Norling and Boye 2013). One of the researchers worked in the organization so researcher bias is likely to be present although this was mitigated where possible by minimizing direct researcher-participant contact. The staff and organization were anonymized to prevent recognition by competitors and peers. Six geologists [P1-P6] were recruited for question 2 and sixteen geophysicists [P7-P22] for question 3. Due to the small sample, face to face engagement with twelve geoscientists [P23-P34] (two groups of six) provided additional qualitative data.

The results from questions 1-3 would be combined with a synthesis of the literature to shape the theoretical model.

Colour coding scheme for methodologies

A colour coding scheme was used to describe the interactions between manual and automated methods used in the research (Figure 2). This colour scheme is also taken into the concluding theoretical model.

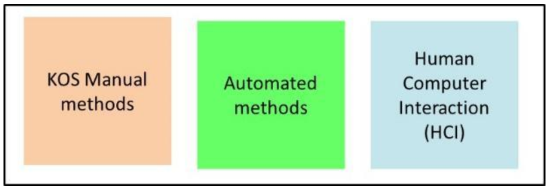


Figure 2 – Colour coding scheme used in the study methodology

Research design for question 1 and 2

Research questions 1 and 2 were addressed through a real business problem identified in the study organization. An oil and gas exploration team in Europe had over 13,000 electronic office documents on their shared file system. These were organized in folders by the team for content navigation and browsing, however the team could only search by filename. This hampered findability as they had new graduates that were not familiar with the folder navigation designs or past individual filing practices. It was commonplace for information to be included as part of a general presentation file that did not have that topic in its filename, to be filed in a folder which also did not have the topic in its folder name. So unless the geoscientist had created the presentation or seen it previously, it was easy to miss this information.

A commercially available enterprise search tool enabled the 13,000 files to be full text indexed. Geoscientists were able to type their queries into a search box and examine results using a web interface. Automatic categorization of content by geoscientific topics (not used in existing folder name/document type) was performed using a commercially available oil and gas thesaurus licensed by the organization (but for the purpose of manually tagging). A geoscience subset of the thesaurus (2,510 concepts) was used for the study including the subject areas of Geological Time, Lithology and Depositional Environments.

The automated categories were presented in a hierarchical faceted search menu on the left hand side of the search User Interface (UI) to provide an overview of search results, containing a bracket and number showing how many documents had been found containing that concept. Actual search results were listed in the middle of the screen. The value of providing a series of visual 'prompts' (the facets) to browse and filter search results would be assessed by geoscientists. The methodological process is shown in Figure 3.

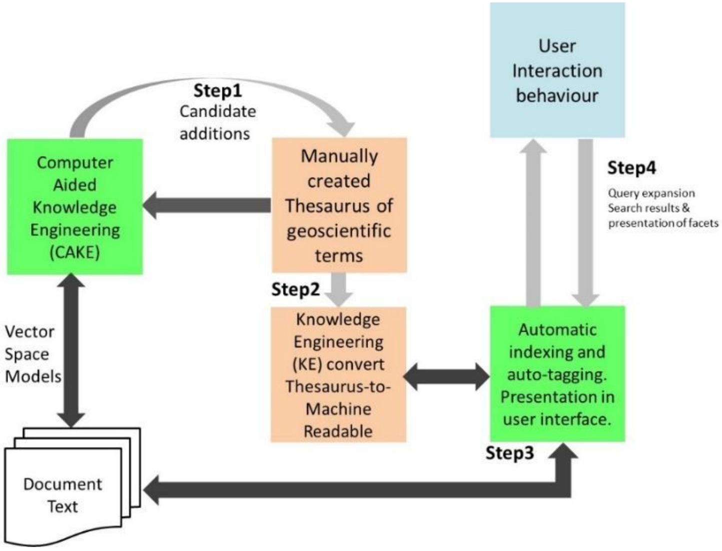

Step1
Candidate additions
Computer Aided Knowledge Engineering (CAKE)
Manually created Thesaurus of geoscientific terms
User Interaction behaviour
Step4
Query expansion
Search results &
presentation of facets
Step2
Vector Space Models
Knowledge Engineering (KE) convert Thesaurus-to-Machine Readable
Automatic indexing and auto-tagging. Presentation in user interface.
Step3
Document Text

Figure 3 – Methodology for research question 1 and 2

Step1 – Computer Aided Knowledge Engineering (CAKE) - Automatic equivalent term identification
Step1 addressed the first research question, to what extent can an oil and gas thesaurus (KOS) be enhanced through automated techniques. In order to focus ultimately on precise AQE (Step3), the goal of identifying additional equivalent terms was chosen to test whether coverage of the existing thesaurus could be enhanced through automated methods. Each concept node and synonym in the thesaurus was automatically compared to a vector space model automatically generated from the text contained in the 13,000 files. The Word2vec algorithm (Mikolov *et al.* 2013) was used, using the top 20 cosine values with a string length greater than five and a Levenshtein edit distance two or less as a cut off for each concept. This is in line with heuristics used in practice elsewhere (Cholakian 2013).

It is not suggested these algorithms are the best performing and it was not part of the research study to compare algorithms and parameters. These methods were deemed sufficient given the research study questions. A random sample of 334 concepts from the geoscience thesaurus was used to evaluate results in order to give 95% confidence in potential term volume increases to the thesaurus (NSS 2014).

Step2 – Converting the thesaurus for automated use – Knowledge Engineering (KE)
Relevant thesaurus associations were made explicit in order to be machine readable, an activity often termed Knowledge Engineering (KE). This was particularly important so transitive associations could be defined for inference. For example, a search query on the concept 'Eocene' would need to expand the query to any equivalents and concept sub-classes (*Priabonian OR Bartonian OR Lutential OR Ypresian).*

Step3 – Use of the modified thesaurus for Automatic Query Expansion (AQE) and Entity Extraction
The thesaurus with the synonym candidate additions (from step 1) and inference rules (from step2), was used to index and categorize the 13,000 documents (not classify to a single document type). Named entity extraction was performed on the text using an existing lookup authority list of 'known' geological basin names for the area concerned as well as automatically generating possible names by extracting all

the nouns that preceded the term *basin* in the text to provide a list of 'possible' basins to present in faceted search as refiners. To improve precision for homonyms, basic intra-domain disambiguation was applied using the concepts from the thesaurus. For example, 'Tertiary' (Geological Time) was disambiguated from Tertiary (Hydrocarbon Migration) and Tertiary (Recovery) by using surrounding terms as 'clues to context'.

Step4 – Gather geoscientist feedback

Step4 addressed the second research question, what is the perceived value of auto-categorizing content that is already manually classified (through folders). The researchers sent the participants a link to the enterprise search tool with some basic instructions and avoided physical contact to minimize observer expectancy bias effects. After a period of two weeks, the participants were sent a semi-structured questionnaire containing four questions via email to gather their feedback.

Unfortunately the research study coincided with unexpected organizational changes that limited participation for this study to only six geoscientists in an exploration team (who volunteered to take part in the study, so are a self-selecting group). The small sample size determined the form of analysis (predominantly qualitative) based on the questionnaire comments, rather than quantitative based from the Likert items in the questionnaire. The questionnaire consisted of the following questions allowing a ranking by Likert scale (*1=not at all*, *5=to a large extent*) and a space for comments:

1. Compared to what exists today, to what extent do these new techniques improve your ability to find information (any why)?
2. If these new techniques were deployed on the file-system what reduction in time spent searching would they make (and why)?
3. To what extent do the new techniques allow you to discover new information and insights that are not possible today (and why)?
4. Rate the most important features (methods) on display in the search User Interface (UI) and why.

Research design for question 3

The methodology in Figure 4 address the third research question, to what extent can manual and automated KOS techniques be combined in a search UI to stimulate serendipity.

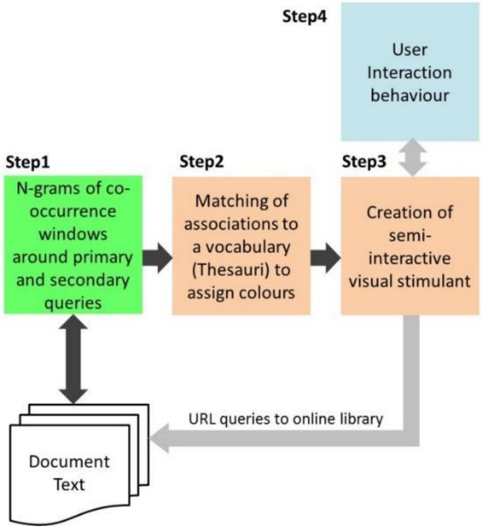

Step4
User Interaction behaviour
Step1
N-grams of co-occurrence windows around primary and secondary queries
Step2
Matching of associations to a vocabulary (Thesauri) to assign colours
Step3
Creation of semi-interactive visual stimulant
URL queries to online library
Document Text

Figure 4 – Methodology to address research question 3

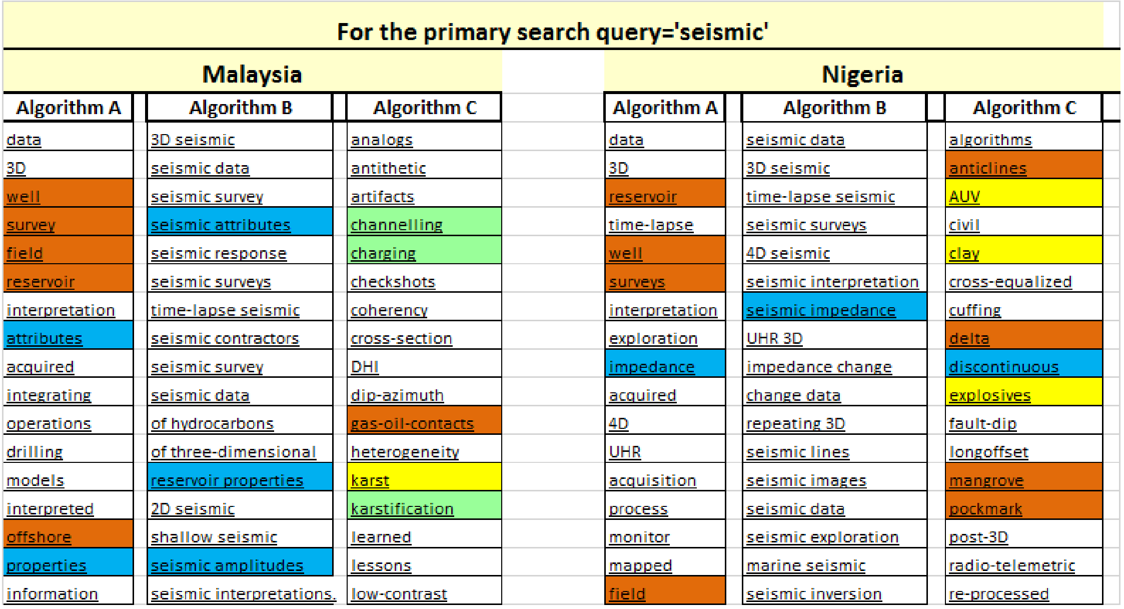
A semi-interactive 'stimulant' was created based on local context. The stimulant was designed to provoke interaction and discussion using content from the Society of Petroleum Engineers (SPE) in the form of 70,000 article abstracts. Sample search queries (primary search query of 'seismic' and secondary country based queries) were chosen from the study organization's search logs to ensure they were representative.

Step1 – Information Extraction (IE)

Python scripts were applied to the 70,000 SPE abstracts, creating co-occurrence networks (using a 16 word window) to the primary search query ('seismic'), where the primary ('seismic') and secondary search terms ('Gulf of Mexico', 'Malaysia', 'Nigeria', 'Australia' or 'Canada') occurred in a 50 word text window. The 50 word window was arrived at deductively by trying smaller and large sizes and examining the number of false positives and false negatives, although the window size is likely to be related to the nature of the specific text collection being analyzed.

Step2 – Colour assignment

The SWEET (Raskin 2011) ontology and the commercial thesaurus (used for questions 1 and 2) was used to colour code the terms through a matching process, to break up the display and highlight potential patterns as shown in (Figure 5).



For the primary search query='seismic'					
Malaysia			Nigeria		
Algorithm A	Algorithm B	Algorithm C	Algorithm A	Algorithm B	Algorithm C
data	3D seismic	analogs	data	seismic data	algorithms
3D	seismic data	antithetic	3D	3D seismic	anticlines
well	seismic survey	artifacts	reservoir	time-lapse seismic	AUV
survey	seismic attributes	channelling	time-lapse	seismic surveys	civil
field	seismic response	charging	well	4D seismic	clay
reservoir	seismic surveys	checkshots	surveys	seismic interpretation	cross-equalized
interpretation	time-lapse seismic	coherency	interpretation	seismic impedance	cuffing
attributes	seismic contractors	cross-section	exploration	UHR 3D	delta
acquired	seismic survey	DHI	impedance	impedance change	discontinuous
integrating	seismic data	dip-azimuth	acquired	change data	explosives
operations	of hydrocarbons	gas-oil-contacts	4D	repeating 3D	fault-dip
drilling	of three-dimensional	heterogeneity	UHR	seismic lines	longoffset
models	reservoir properties	karst	acquisition	seismic images	mangrove
interpreted	2D seismic	karstification	process	seismic data	pockmark
offshore	shallow seismic	learned	monitor	seismic exploration	post-3D
properties	seismic amplitudes	lessons	mapped	marine seismic	radio-telemetric
information	seismic interpretations.	low-contrast	field	seismic inversion	re-processed

Figure 5 – Part of the semi-interactive stimulant for the primary query 'seismic' and secondary query 'Malaysia' and 'Nigeria'. Presenting co-occurrence associations colour coded by KOS (*orange=realm, green=natural process/phenomena, yellow=matter/materials, blue=property*)

Step3 – Creation of semi-interactive stimulant

The representation is shown as a list as opposed to a three dimensional representation because more terms can be included and people may find vertical lists faster to scan than representations that display terms from left to right (Halvey and Keane 2007). Fifty associations (rows) were displayed to increase the chances of serendipitous encounters, with evidence scientists find interesting associations outside the top ten or twenty typically shown in faceted enterprise search menu's (Cleverley and Burnett 2015a).

Algorithm A displayed representative unigrams and Algorithm B bigrams (both ranked by descending frequency of occurrence). Algorithm C was a discriminatory set of words (unique clues for that context combination) ranked alphabetically. An example for the latter is 'Karst' for 'Malaysia', indicating this term co-occurs with 'Seismic' and 'Malaysia' but not for 'seismic' and any of the other secondary search terms. It is not suggested these are the optimal algorithms to produce 'surprising' associations but were used because they each delivered significantly different terms with respect to specificity and descriptiveness. The extent to which Algorithms A, B or C could stimulate unexpected associations would be investigated. Each cell was linked via URL's to the SPE online digital library, allowing staff to click through and see the search results and documents *in which that association occurs locally*.

Step4 – Researching user interaction

Sixteen geophysicists took part in the use case study, company staff were purposefully sampled (Coyne 1997) to ensure representation from every geophysical department in the study organization. Focus groups (Morgan 1997) were used to enable the researchers to quickly identify a full range of perspectives held by respondents, "*the interactional, synergistic nature of the focus group allows participants to clarify or expand upon their contributions to the discussion in the light of points raised by* *other participants…that might be left underdeveloped in an in-depth interview"* (Powell and Single 1996, pg. 504). Each session lasted 45mins and consisted of between two to nine staff. The stimulant (Figure 6) was made available on large touchscreens at the organization's premises and participant interactions were video-recorded.

Figure 6 – The semi-interactive stimulant on the large touchscreen with participant interactions

Results

The results for questions 1, 2 and 3 are provided with discussion. The findings are combined with the literature leading to the formation of a theoretical model presented in the next section.

Q1. To what extent can an oil and gas thesaurus be enhanced through automated techniques?

An analysis of the candidate terms produced by statistical vector space models for 334 random concepts in the thesaurus yielded the following data (Table 1).

Table 1 – Example equivalent terms identified by statistical techniques from the 13,000 files

Example Type	**Automatically extracted equivalence terms are in brackets**
Lexemes	Vitrinite (*Vitrinites*), Tuff (*Tuffaceous*), Cataclasite (*Cataclasitic*)
New synonyms	Rhyolite (*Metarhyolite*), Monzonite (*Monzogranite*)
Spacing/Spelling	Clay shale (*Clayshale*), Wackestone (*Wackstone*)

The random sample of 334 concepts generated a 34% increase in valid thesaurus terms using this approach. This is considered significant, based on the size and depth of the existing commercial thesaurus where subject matter experts had already explicitly modeled synonyms and lexemes.

Errors were also identified, for example *volcanic ash* suggested as a synonym for *volcanic glass*, they are associated but not the same. There may be value for corporate taxonomists to use these types of statistical techniques as a best practice to augment their modeling efforts particularly for synonyms and lexemes as its unlikely all combinations and variants can be modelled manually by an individual or small group. This supports existing research on the value of CAKE methods (Salmador-Sanchez and Angeles-Palacios 2008) and could be taken further to create a first pass associative network (where one does not exist) to be refined manually by subject matter experts. The results provide evidence that combining manual and automated techniques in a 'mixed methods' approach on the same KOS, delivers a level of quality that a single method (manual OR automatic) is unlikely to deliver.

Q2. What is the perceived value of auto-categorizing content that is already manually classified using folders?

There was unanimous agreement from participants that having the ability to search documents full text (rather than just by filename) and browse auto-categorized facets (rather than just folder classifications), improved the ability to find and discover information *to a large extent*. The average number of tags added automatically (from the KOS counting unique leaf values only) was 113.9 per document for Adobe PDF files and 23.25 for other office files. As part of the iterative process of inquiry, reports were run on two random EDMS (SharePoint) areas consisting of 103 documents added by ten geoscientists in the study organization, yielding an average of 3.6 tags per document (where two mandatory pick lists were in operation) and 1.1 tags per document (with a single optional pick-list). This illustrates the potential value of auto-categorization in complimenting manual tags, delivering a 'richness' of tags on topics for faceted search menu's to support browsing, that manual tagging is unlikely to deliver.

The mode of questionnaire responses for time saved (through the new techniques) was 50%. Business value based themes identified included increased productivity and discoverability, evidenced respectively by *"Reports hidden in the system where no-one could find them. To search in all these folders, often titles don't describe enough what information they hold, it takes weeks. This system takes seconds!! Time saved is unmeasurable".* [P2] and *"The search is more thorough….and allows you to put your search word in context. You can find resources you may never have come across otherwise."* [P6]. A theme of improving quality was identified, "*the current system of not being able to find documents encourages people to save multiple copies in different directories. This could help reduce duplication."* [P3].

The full text search ('Google like' search window) and faceted search menu were rated equally as important by participants. This finding may contradict research reporting low usage of faceted search (Ballard and Blaine 2011, Niu and Hemminger 2010). The search mode of participants in this study (*"The majority of our searching is exploratory!"* [P3]), and detailed nature of the facet values (richer than that which is likely to be added manually) may explain differences. Being prompted with facets in advance, showing what is contained in a corpus or within search results was considered advantageous, *"The fact that the tool provides the user with keywords, it reduces the time to think about the keywords"* [P5].

Despite the problems caused by folders, there was a strong desire from all participants to keep folder structures as a means to classify manually and find it back when they knew what they were looking for. All participants were also keen to have an option to link from any file they found in a search result page using the full text search and/or faceted refiners, to the folder in which it was located. One reason given was, *"9 times out of 10 (it would help) to find other critical information"* [P3].

Some geological basin entity names found automatically through extraction of the nouns preceding the word 'basin' in text (and presented as refiners) which were considered useful by participants, were not on the authority list of basin names. This supports Greenberg (2011), KOS *can* limit new discoveries.

Deeper issues regarding search in the enterprise were revealed within the themes of file permissions, information behaviours and search literacy. For example, *"Often the 'hidden gems' that you accidently come across are in confidential folders"*, [P4]. *"Great concept. Obviously, it will work even better if a culture of adding good keywords to all documents can be implemented."* [P4]. During subsequent engagements with twelve geoscientists, it was clear that most were not aware of the role semantics plays in exploratory searching. For example, a query on *limestone play* will not return items on *oolite play* that do not mention *limestone.* A geologist knows *oolite* 'is-a' *limestone,* but a keyword search engine does not. This can be summarized in the comment, *"I learnt that Google is not a Geologist"* [P23]. This may have implications for search literacy leading to potentially sub-optimal search task outcomes.

Q3. To what extent can manual and automated KOS techniques can be combined in a search user interface to stimulate serendipity?

Serendipitous encounters were documented during interactions with the stimulant, including:

> *"Word associations highlighted new and unexpected terms such as 'metamorphic sole' associated with the secondary keyword 'platform'. This surprising result led us to consider a new geological element which could impact our (exploration) opportunity"* [P32].

Fifteen of the sixteen participants (94%) thought the use of colour to classify associations aided pattern identification. This provides a possible example where combining manual and automated KOS/KO techniques together in a 'mixed methods' approach on the *same* collection of content, delivers value which a single method (manual OR automatic) is unlikely to deliver. Other themes that emerged:

Differing intents and behaviours

On seeing the stimulant for the first time, participants commented both, "*This is overwhelming – too much"* [P11] and *"Excitement is the first thought I had"* [P28]. From a KM perspective, individual behaviours and personal preferences may affect the take up and exploitation of these techniques.

Help manage taxonomy creation

Using the technique to help create geophysical taxonomies was identified as a potential use, "*could be extremely useful in the debate that is unravelling about the taxonomy, because taxonomy is difficult"* [P9],"*Yes this could help as a data driven taxonomy, very powerful"* [P7]. It appears that the geophysicists had not thought of using automated techniques as input to their manual taxonomy modeling work.

Some algorithms are more likely to generate unexpected and serendipitous encounters than others
All participant's preferred Algorithm C, over Algorithms A and B. For example, "*some of them attract my attention because they are very unique, most is not unique (e.g. seismic mapping) these are categories. I am looking for unique things that trigger my attention this would be a starting point".* [P12]. Mismatches between the searchers mental model and stimulant associations, "*It is like open up the box for me and I pick what does not fit with my brain, like one of those games"* [P14] triggered interest from participants.

Geoscientists like to browse and navigate concepts, do not always know what search terms to use
As in the data from question 2, participants found the ability to browse filters useful, acting as prompts to make searches they may have otherwise not made, *"This helps with big problem with Google (or that I have with Google), is choosing right selection of words to find something.."* [P13].

Theoretical Model

A theoretical model (Figure 7) is presented in fulfillment of the research objective. The KO/KOS methods labelled 1-7 in Figure 7 are explained in Table 2, tied to the three research questions (Q1, Q2 and Q3).

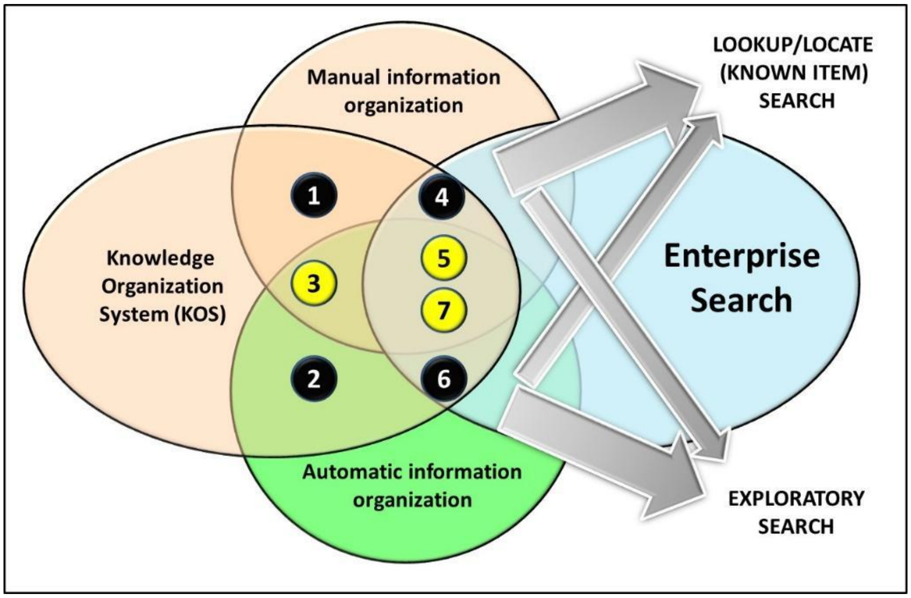

LOOKUP/LOCATE (KNOWN ITEM) SEARCH
Manual information organization
Knowledge Organization System (KOS)
1
2
3
4
5
6
7
Enterprise Search
Automatic information organization
EXPLORATORY SEARCH

Figure 7 – Theoretical model showing the different KO approaches (manual and automatic). The wider the arrow the greater the emphasis for that method supporting the respective search goal.

The mixed methods approaches derived from the primary data from this research (methods 3, 5 and 7) are shown in yellow in Fig. 7 and Table 2. The other methods are supported by the existing literature. The concepts of *multi-method* and *mixed methods* (usually used to describe research methods) is applied analogously to the application of automated and manual KO/KOS methods in the enterprise. In this context, the literature review supports using a *multi methods* strategy in the enterprise with respect to KOS/KO (automatic AND manual), not a single method (manual OR automatic) for different content and search goals. The primary data from this research, supported by the literature, also provides examples where combining manual/automated methods together *(mixed-methods)* on the *same* collection of content or KOS, provides synergies which are likely to exceed those of a single method.

Manual KO methods appear to *predominantly* support Lookup search, whilst automated methods appear to *predominantly* support exploratory search. A *pluralist* enterprise KO strategy (manual and automated KOS/KO) is likely to produce better search & discovery outcomes than a *totalitarian* single approach.

Table 2 – KOS/KO approaches from Fig. 7 for manual and automated methods. Methods 3, 5 and 7 (in Yellow) have been derived from the primary research data in questions (Q1, Q2 and Q3).

	NO.	METHOD	DESCRIPTION
KOSStructureCreationMethod	1	**Single method** (Manually developed KOS)	As stated by Velardi *et al.* (2012), it is virtually impossible to re-create domain specific complex taxonomies ('is-a' or 'part-of' associations) automatically from text. Manual taxonomies and authority lists mainly support precise known item search (Marchionini 2006).
	2	**Single method** (Automatically developed KOS)	Manually generated KOS are time consuming to create and may limit new discoveries (Greenberg 2011). Unsupervised machine learning can quickly create topics and associations (Palmer *et al.* 2001) and can be applied to content *independent* to that which created it.
	3	**Mixed method** (Manual and Automatic) developed KOS	The results from this study (**Q1**) provide evidence for how semi-supervised statistical techniques applied to enterprise content, can enhance the quality of a manually developed KOS by adding 34% more synonymous (equivalence) terms. The conclusion is that automated techniques can enhance manual KOS creation to a large extent.

KOMethod	4	**Single method** (Manual classification and/or categorization	For precise searching, manual organization is required, especially when it is not possible for automated techniques to infer the class, provenance or importance of information (e.g. clues may not be available in the text). For example, stage gate business deliverables (Abel and Cleverley 2007) & lessons learned (Piantanida *et al.* 2015).
	5	**Mixed method** (Automatic classification and/or categorization (to a KOS)	The results from this study (**Q2**) provide evidence for how auto-categorization techniques using a manually created KOS can enhance search recall & information discovery to a large extent, even if the content has already been manually classified or categorized. Value in applying to large volumes of unclassified and diverse information (e.g. email/news) which is too costly/time consuming for manual methods.
	6	**Single method** (Automatic organization)	Unsupervised machine learning techniques can organize information quickly and cheaply, without the need for existing expensive KOS. These techniques may surface real world associations that would not be discovered by using a KOS (Yu *et al.* 2014, Blei *et al.* 2003).
	7	**Mixed method** (Manual and Automatic) organization)	The results from this study (**Q3**) provide evidence for how a manually created KOS can be applied automatically to colour unsupervised word co-occurrence clustering to facilitate serendipitous information discoveries. Use of other methods provide areas for further research.

Conclusion

The development of Knowledge Organization Systems has evolved through different disciplines over time. The clear separations that may have existed in the past between Library and Information Science, Data Management, KM, BI, IR and Artificial Intelligence have converged and even overlapped.

As part of a strategic approach, there is a good case to adopt *multiple methods* and use manual and automated KO and KOS approaches for different types of oil and gas content, underpinned by good governance. In addition, it has been shown there are synergistic benefits to using *mixed methods* approaches (blending manual and automated approaches together) applied to the *same* collection of content or KOS. Enhancing KOS quality (with resulting findability benefits) and increasing the propensity of a search UI to facilitate unexpected, insightful and serendipitous discoveries are two such benefits. It is proposed that these synergies are likely to deliver outcomes not possible using a single approach, improving search and discovery performance in the enterprise. The theoretical model proposed may help re-conceptualize understanding in this area and provide input into KM, IM and IT strategies.

Practical applications of the research may exist in two areas. Firstly, an organization could evaluate their current information search & discovery and classification practices using the theoretical model, which may present opportunities for improvement. This may range from leveraging existing content more effectively, through to introducing new practices and possibly new technologies based on the premise that it is becoming increasingly challenging to read all relevant information. Secondly, an organization could ensure their information professionals are *multilingual* in the language of all the disciplines that interact with KOS on the basis that innovation often happens at these functional boundaries. Embracing established and emerging computer science techniques is one such discipline. This holistic approach could increase the corporate information professionals' ability to *proactively* stimulate business needs and opportunities, not just react to them.

The development and potential transferability of the theoretical model to other industry sectors offers further areas for research.

Acknowledgements

Appreciation is given to the company that acted as the research case study for their support. Thank you to John, Raji and Chao for their crucial help support and expertise. Appreciation is also given to the Society of Petroleum Engineers (SPE) who allowed their information to be used for this research. A final thank you to the anonymous reviewers for their input and guidance.

References

Abel, R., Cleverley, P. H. (2007). Improving Information Delivery. Hart's E&P March Edition. Online Article (Accessed February 2013).

Adkins, S (2003). Information Gathering in the Electronic Age: The Hidden Cost of the Hunt. Safari Techbooks, January 2003.

AIIM (2008). Market IQ Report: Findability: The art and science of making content easy to find. Association for Information and Image Management (AIIM) 2008. Sponsored by OpenText.

Allan, J., Croft, B., Moffat, A., Sanderson, M. (2012). Frontiers, Challenges, and Opportunities for Information Retrieval. Report from the Second Strategic Workshop on Information Retrieval in Lorne, February 2012, ACM SIGIR Forum, 46(1), 2-32

Alyahyaee, A. (2012). Oil & Gas Data Repository (OGDR), Energistics National Data Repository (NDR) '11 Update. 21st-24th October 2012. Kuala Lumpur, Malaysia.

Andersen, E. (2012). Making Enterprise Search Work: From Simple Search Box to Big Data Navigation. Center for Information Systems Research (CISR) Massachusetts Institute of Technology (MIT) Sloan School Management, 12(11).

Ballard, T., Blaine, A. (2011). User search limiting behaviour in Online Catalogs. Comparing classic catalog use to search behaviour in next generation catalogs. New Library World, 112(5/6), 261-273.

Bawden, D. (1986). Information-Systems and the Stimulation of Creativity. Journal of Information Science, 12(5), 203-216.

Behounek, S., Casey, K. (2007). EarthSearch=GoogleEarth Enterprise+PetroSearch. Society of Petroleum Engineers (SPE) Digital Energy Conference and Exhibition, 11-12th April, Houston, Texas, USA. Report ID: SPE-108208-MS

Berger, P.L., Luckmann, T. (1966). The social construction of reality. A treatise in the sociology of knowledge. 1st ed. London: Penguin.

Bhogal, J., Macfarlane, A., Smith, P. (2007). A review of ontology based query expansion. Information Processing and Management, 43, 866-886.

Blackman, S. (2012). Risky business: challenges of deepwater drilling in the North Sea. Offshore Technology, 21st June 2012. Online Article (Accessed December 2014).

Blei, D, Ng, A., Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 2003*,* 3, 993-1022

Caballero, R, Nuernberg, S. (2014). Building an Enterprise Taxonomy. 18th International Petroleum Data, Integration and Data Management (PNEC), May 20-22nd 2014, Houston, USA.

Carpineto, C., Romano, G. (2012). A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys, 44(1), 1-50.

Cholakian, A. (2013). How to Use Fuzzy Searches in Elastisearch. Online Article (https://www.found.no/foundation/fuzzy-search/, accessed January 2015).

Chuang, J., Manning, C.D., Heer, J. (2012). "Without the Clutter of Unimportant Words": Descriptive Keyphrases for Text Visualization. ACM Transactions on Computer-Human Transactions, 19(3)

Chui, M., Manyika, J., Bughin, J., Dobbs, R., Roxburgh, C., Sarrazin, H., Sands, G., Westergren, M. (2012). The social economy: Unlocking value and productivity through social technologies. McKinsey Global Institute Report. Online Article (Accessed January 2015).

Chum, F., Everett, M., Hills, S., Soma, R., Cutler, R. (2011). Realizing the Semantic Web Promise in the Oil & Gas Industry: Challenges and Experiences. SemTech 2011,, 9th June 2011, San Francisco, USA.

Cleverley, P.H. (2012). Improving Enterprise Search in the Upstream Oil and Gas Industry by Automatic Query Expansion using a Non-Probabilistic Knowledge Representation. International Journal of Applied Information Systems (IJAIS), 1(1), 25-32

Cleverley, P.H., Burnett, S. (2015b). Creating sparks: comparing search results using discriminatory search term word co-occurrence to facilitate serendipity in the enterprise. Journal of Information and Knowledge Management (JIKM).

Cleverley, P.H., Burnett, S. (2015a). Retrieving haystacks: a data driven information needs model for faceted search. Journal of Information Science, 41, 97-113

Collins, J.C., Porras, J.I. (1997). Built to Last. Successful Habits of Visionary Companies. New York, HarperCollins.

Coyne, I.T. (1997). Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries. Journal of Advanced Nursing, 26, 623-630.

Dale, E. (2013). The importance of constant measurement in search relevance. A longitudinal case study. Ernst & Young. Enterprise Search Summit 2013, New York, USA.

DeLone, W.H., McLean, E.R. (2002). The DeLone and McLean Model of Information System Success: A Ten Year Update. Journal of Management Information Systems, 19(4), 9-30.

Delphi (2002). Taxonomy & Content Classification. Market Milestone Report. Online Article (Accessed March 2013).

Demartini, G. (2007). Leveraging Semantic Technologies for Enterprise Search, PIKM November 2009, Lisboa, Portugal.

Dextre Clarke, S.G., Zeng, M.L. (2012). From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling. Information Standards Quarterly, 24(1), 20-26.

Doane, M. (2010). Cost benefit analysis: Integrating an enterprise taxonomy into a SharePoint environment. Journal of Digital Asset Management, 6(5), 262-278

Fagan, J.C. (2010). Usability studies of faceted browsing: A literature review. Information Technology and Libraries, 58-66.

Faith, A. (2011). Linguistically Training Automatic Indexing Software for Complex Taxonomies. Semantic Technology & Business Conference June 2013.

Feldman, S., Sherman, C. (2001). The High cost of not finding information. White Paper International Data Corporation (IDC).

Feldman, S., Marobella, J.R., Duhl, J., Crawford, A. (2005). The Hidden Costs of Information Work. White Paper International Data Corporation (IDC).

Feldman, S. (2009). IDC Executive Briefings: Information Advantage: Information Access in Tommorow's Enterprise. International Data Corporation (IDC).

Foster, A. & Ford, N. (2003) Serendipity and information seeking: an empirical study. Journal of Documentation. 59(3), 321-340

Friedman, B. (2010). Serendipity is an Explorationists best friend. American Association of Petroleum Geologists (AAPG) Online Article.

Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 964-971

Garbarini, M., Catron, R.E., Pugh, B. (2008). Improvements in the Management of Structured and Unstructured Data. Society of Petroleum Engineers, Report IPTC12035.

Ghiselin, D. (2010). Serendipity is alive and well at EagleFord. Hart's E&P Online Article.

Gimmal (2013). Information Governance and Compliance in Oil and Natural Gas Company. Online Article (Accessed January 2015)

Goker, A., Davies, J. (2009). Information Retrieval: Searching in the 21st Century. UK: Wiley & Sons Ltd

Greenberg, J. (2011). Introduction: Knowledge Organization Innovation: Design and Frameworks. Bulletin of the American Society for Information Science and Technology, April/May 2011, 37(4), 12-14.

Grefenstette, G. (1994). Explorations in Automatic Thesaurus Generation. MA, USA: Kluwer Academic Publishers Norwell

Grimes, S. (2014). Text Analytics Applied. 2nd LIDER Road mapping workshop, May 8th 2014, Madrid, Spain.

Gwizdka, J. (2009). What a difference a tag cloud makes: effects of tasks and cognitive abilities on search results interface use. Information Research. 14(4)

Halvey, M., Keane, M.T. (2007). An assessment of tag presentation techniques. Proceedings of 16th International World Wide Web Conference (WWW).

Hamski, J. (2010). Unstructured Geospatial Information for a Competitive Advantage in Resource Exploration. Elsevier, Online Article, Accessed January 2015.

Hearst, M.A. and Stoica, E. (2009). NLP Support for Faceted Search Navigation in Scholarly Collections. Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP Suntec, Singapore 7th August 2009, 62-70

Hedden, H. (2013). Taxonomies for Auto-Tagging Unstructured Content. Text Analytics World, October 1st 2013, Boston USA.

Heye, D. (2003). Taxonomies and automatic classification at Shell – a case study. 'Building a Knowledge Framework: Practical Taxonomy Design and Application Conference, September 29-30th Amsterdam, The Netherlands.

Hills, S. (2014). Why we Want to Implement ISO Metadata: Energy Industry Profile of ISO 19115-1:2014 ("EIP") V1.0. Energistics FGDC ISO Metadata Implementation Forum 12th February 2014.

Hjorland, B. (2008). What is Knowledge Organization (KO)? International journal devoted to concept theory, classification, indexing and knowledge representation, 35(2/3), 86-101

Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Washington, USA, First Digital Library Federation and Council on Library and Information Resources.

Hubert, C. (2012). Seamless Collaboration. Enabling Employees to Work Together Across Boundaries. APQC Report K03906, 1-15.

Jacob, E.J. (2004). Classification and categorization: A Difference that Makes a Difference. Library Trends, 52(3), 515-540.

Jacobs, P.S., Rau, L.R. (1990). SCISOR: Extracting information from on-line news. CACM 33, 88-97

Jurka, T.P., Collingwood, L., Boydstun, A.E., Grossman, E., van Atteveldt, W. (2013). RTextTools: A Supervisory Learning Package for Text Classification. The R Journal, 5(1), 6-12.

Kaizer, J., Hodge, A. (2005): "AquaBrowser Library: Search, Discover, Refine", Library Hi Tech News, 22(10), 9-12

Khoo, C.S.G., Luyt, B., Ee, C., Osman, J., Lim, H., Yong, S. (2007). How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour. Information Research, 11(2).

Krestel, R., Demartini, G., Herder, E. (2011). Visual Interfaces for Stimulating Exploratory Search. JCDL 2011, June 13th-17th Ottawa, Canada, 393-394.

Landauer, T.K., Dumais, S.T. (1997). A Solution to Platos' Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104(2), 211-240.

Lennon, A., Alshubi, F., Cleverley, P.H. (2012). Improving Subsurface and Wells Document Management at Qatar Shell. 16th Annual Petroleum Data Integration Conference. May 15th-17th Houston, USA.

Low, B. (2011). Usability and contemporary user experiences in digital libraries. CIGS Seminar, University of Edinburgh. Slide 17

Lowe, A., McMahon, C., Culley, S. (2004). Characterising the requirements of engineering information systems. International Journal of Information Management, 24, 401-422.

Luke, T., Schaer, P., Mayr, P. (2012). Improving Retrieval Results with discipline-specific Query Expansion. Proceedings of Theory and Practice of Digital Libraries, 2012.

Lykke, M., Eslau, A.G. (2010). Using Thesauri in Enterprise Settings: Indexing or Query Expansion? The Janus Faced Scholar: a Festschrift in honour of Peter Ingwersen. Det Informationsvidenskabelige Akademi, 87-97

Magnuson, D. (2014). Auto Classification and the Holy Grail for Records Managers. IBM Presentation as the Association or Records Managers and Administrators (ARMA), Houston.

Majid, S., Anwar, M.A., Eisenshitz, T.S. (2000). Information Need and Information Seeking Behavior of Agricultural Scientists in Malaysia. Library & Information Science Research, 22 (2), 145-163

Manning, C.D., Schutze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, United States of America, Massachusetts Institute of Technology (MIT) Press.

Manning, C.D., Raghavan, P., Schutze, H. (2009). An Introduction to Information Retrieval. Cambridge, England. Cambridge University Press.

Marchionini, G. (2006). Exploratory Search: From Finding to Understanding. Communications of the ACM. 49 (4), 41-46

Martela, F. (2015). Fallible Inquiry with Ethical End-in-View: A Pragmatist Philosophy of Science for Organizational Research. Organizational Studies, 1-27.

McCandless, D. (2012). Information in beautiful, 2nd ed., William Collins, London.

McCay-Peet, L. & Toms, E. (2011). Measuring the dimensions of serendipity in digital environments. Information Research, 16(3)

McDonald, S., Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. Proceedings of the 23rd Annual Conference of the Cognitive Science Society, 611-616.

McNaughton, N. (2015). Knowledge organization – the great debate! Oil Information Technology Journal, 20(2), 1-11

Meza, D. (2014). On Developing Better Magnets for Needles in Haystacks. Office of the Chief Knowledge Officer (CKO), National Aeronautical Space Administration (NASA). NASA Online Article and Interview, Accessed December 2014.

Microsoft and Accenture (2010). Upstream Oil & Gas Computing Trends Survey (2010). Conducted by PennEnergy Research and the Oil & Gas Journal Research Centre.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advanced in Neural Information Processing Systems, 3111-3119

Miller, D. (2014). Just the facts: Auto-classification and Taxonomies. ConceptSearching Webinar, Online Article (Accessed February 2015).

Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K. (1990). WordNet: An online lexical database. International Journal of Lexicography, 3(4), 235–244

Mindmeter (2011). Mind the Enterprise Search Gap. Report Sponsored by SmartLogic.

Morgan, D.L. (1997). Focus Groups as Qualitative Research: Planning and Research Design for Focus Groups. In Sage Research Methods, 32-46

Morville, P., Rosenfeld, L. (2006). Information Architecture for the World Wide Web: Designing Large-Scale Websites. 3rd Edition, O'Reilly.

Munkvold, B.E., Paivarinta, T., Hodne, A.K., Stangeland, E. (2006). Contemporary issues of enterprise content management: the case of Statoil. Scandinavian Journal of Information Systems, 18(2), 69-100.

Navigli, R., Velardi, P. (2002). Automatic Adaptation of WordNet to Domains. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC '02), Canary Islands, Spain.

Niu, X., Hemminger, B.M. (2010). Beyond Text Querying and Ranking List: How People are searching through Faceted Catalogs in Two Library Environments. Proceedings of the 73rd Association for Information Science and Technology (ASIS&T) Annual Meeting on Navigating Streams in an Information Ecosystem 2010, 47(29)

Noor, A.M., Yassin, C.Z.H. (2006). Issues, Challenges and Constraints in K-Era. Proceedings of the Knowledge Management International Conference. Kuala Lumpur, Malaysia, 6-8th June 2006.

Norling, K., Boye, J. (2013). 2013 Findability Survey. Findability Day. Findwise, Stokholm May 2013

NSS (2014). National Statistics Service Australia Online Calculator (Accessed September 2014).

O'Donnell, M. (2011). Visualizing Patterns in Text: Keynote talk at AESLA (Spanish Association of Applied Linguistics), University of Salamanca May 4th-6th. (Online Article, accessed September 2014).

Ohly, P.H. (2012). Actas del X Congreso ISKO Capitulo Espanol (Ferrol 2012), 541-551

Oil and Gas UK (2011). Oil and Gas UK. Exploration Economic Report 2011. Online Article (Accessed January 2015).

Olson, T.A. (2007). Utility of a faceted catalog for scholarly research. Library Hi Tech. 25(4), 550-561.

Oracle (2012). From overload to impact: An industry scorecard on big data business challenges. Online Article (Accessed March 2013).

Outsell (2005). Survey of Knowledge Workers. Online Article (Accessed March 2013).

Painter, K., Dutton, S.J., Owens, E.O., Burgoon, L.D. (2014). Automatic Document Classification for Environmental Risk Assessment. PeerJ PrePrints,

Palkowsky,B. (2005). A New Approach to Information Discovery – Geography Really Does Matter. Society of Petroleum Engineers (SPE) Annual Technical Conference and Exhibition, Dallas, Texas, USA, 9-12th October 2015. Report ID: SPE 96771

Palmer, C.R., Pesenti, J., Valdes-Perez, R.E., Christel, M.G., Hauptmann, A.G., Ng, D., Wactlar, H.D. (2001). Demonstration of hierarchical document clustering of digital library retrieval results. Proceedings of the 1st ACM/IEEE-CS joint conference on digital libraries, 451.

Peng, J., He, B., Ounis, I. (2009). Predicting the Usefulness of Collection Enrichment for Enterprise Search. ICTIR 2009, 366-370.

Piantanida, M., Cheli, E., Gheorghiso, O., Rossi, P. (2015). Processes and Tools to Effectively Leverage on Lessons Learned for E&P Development Projects. Offshore Mediterranean Conference and Exhibition, 25-27th March, Ravenna, Italy.

Powell, R.A., Single, H.M. (1996). Methodology Matters – V. International Journal for Quality in Health Care, *(5), 499-504

Preece, A., Flett, A., Sleeman, D., Curry, D., Meany, N., Perry, P. (2001). Better Knowledge Management through Knowledge Engineering. Knowledge Management IEEE Intelligent Systems, Jan/Feb 2001, 36-42

Prince, V., Roche, M. (2009). Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration. New York, USA, Medical Information Science Reference.

Quaadgras, A., Beath, C.M. (2011). Leveraging unstructured data to capture business value. Center for Information Systems Research (CISR). MIT, Sloan School of Management, 11(4).

Raskin, R. (2011). National Aeronautical Space Administration (NASA) Semantic Web for Earth and Environmental Terminology (SWEET) Ontology.

Rasmus, D.W. (2013). How IT Professionals can Embrace the Serendipity Economy. Harvard Business Review, August 19th 2013 Online Article (Accessed January 2013).

Robinson, M.A (2010). An empirical analysis of engineer's information behaviors. Journal of the American Society for Information Science and Technology, 61(4), 640-658

Roitblat, H.L., Kershaw, A., Oot, P. (2009). Document categorization in legal electronic discovery: computer classification vs. manual review. Journal of the Association for Information Science and Technology, 61(1), 70-80

Romero, L. (2013). Deloitte: Improving Findability in the Enterprise. APQC Knowledge Management Conference May 3rd 2013, Houston, Texas, USA.

Rose, D.G. (2010). Apache Corporation. The ECM Journey. AIIM Southwest Chapter, May 6th 2010.

Salmador Sanchez, M.P., Angeles Palacios, A. (2008). Knowledge-based manufacturing enterprises: evidence from a case study. Journal of Manufacturing Technology Management, 19(4), 447-468.

Salthe, S.N. (2012). Hierarchical Structures. Axiomathes, 22, 355-383

Sasaki, Y. (2008). Automatic Text Classification. University of Manchester. Online Article (Accessed November 2014).

Schlumberger (2008). Schlumberger Oilfield glossary. Online resource (accessed March 2014).

Shiri, A.A., Revie, C.W., Chowdhury, G. (2002). Thesaurus-assisted search term selection and query expansion: a review of user-centred studies. Knowledge Organization, 29(1), 1-19.

Sidahmed, M., Coley, C.J., Shirzadi, S. (2015). Augmenting Operations Monitoring my Mining Unstructured Drilling Reports. Society of Petroleum Engineers (SPE), SPE-173429-MS.

Skoglund, M., Runeson, P. (2009). Reference-based search strategies in systematic reviews. Proceedings of the 13th International Conference on Evaluation and Assessment in Software Engineering (EASE). Durham University, 20-21st April 2009, 31-40.

Smiraglia, R.P., van den Heuvel, C. (2011). Idea Collider: From a Theory of Knowledge Organization to a Theory of Knowledge Interaction. Bulletin of the American Society for Information Science and Technology, April/May 2011, 37(4), 43-47.

Smith, R. (2012). Implementing Enterprise Information Management at Marathon Oil. Gartner Portals, Content and Collaboration Summit. Track B: Content and Information Management Session B2, March 12th 2012.

Solskinnsbakk, G., Gulla, J.A. (2008). Ontological Profiles as Semantic Domain Representations. NLDB 2008, LNCS 5039, pg. 67-78

Spiteri, L.F. (2004). Word Association Testing and Thesaurus Construction. Library and Information Science Research Electronic Journal (LIBRES), 14(2)

Stenmark, D. (2008). Identifying clusters of user behaviour in Intranet Search Engine log files. Journal of the American Society for Information Science and Technology, 59(14), 2232-2243.

Stock, W.G. (2010). Concepts and Semantic Relations in Information Science. Journal of the American Society for Information Science and Technology, 61(10), 1951-1969.

Strauss, A. & Corbin, J.A. (1998). Basics of Qualitative Research. Techniques and Procedures for Developing Grounded Theory. 2nd Edition Sage Publications.

Tonstad, K., Bjorge, E. (2003). Data Management Metrics in Statoil, Smi Data Management Presentation, London, UK.

Tudhope, D., Alani, H., Jones, C. (2001). Augmenting Thesaurus Relationships: Possibilities for Retrieval. Journal of Digital Information (JODI), 1(8).

Velardi, P., Navigli, R., Martinez, S. (2012). A New Method for Evaluating Automatically Learned Terminological Taxonomies. Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC 2012), May 21-27th, 2012.

Villena-Roman, J., Collada-Perez, S., Lana-Serrano, S., Gonzalez-Cristobal, J.C. (2011). Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 323-328.

W3C (2009). W3C workshop on Semantic Web in Oil and Gas Industry – Report.

Walkup, G.W., Ligon, B.J. (2006). The Good, Bad and Ugly of Stage-Gate Project Management Process as Applied in the Oil and Gas Industry. Society of Petroleum Engineers (SPE) Annual Technical Conference and Exhibition, 24-27th September, San Antonio, Texas, USA. Report ID: SPE-102926-MS.

Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M.X., Qian, W., Shi, L., Tan, L., Zhang, Q. (2010). TIARA: A Visual Exploratory Text Analytic System. Proceedings of ACM. Knowledge Discovery in Databases (KDD), July 25-28th Washington DC, USA.

Wessely, J. (2011). Text Analytics and Auto-Categorization in Semantic Web Applications. SemTech 2011. Online Presentation (Accessed December 2014).

White, M. (2012). Enterprise Search. 1st Edition. California: O'Reilly.

White, M. (2014). Search Strategy A-Z List of Topics. Intranet Focus, September 2014, Online Article.

Wilson, T.D. (2000). Human Information Behavior. Special Issue on Information Science Research, Informing Science, 3(2)

Yu, K., Zhang, J., Chen, M., Xu, X., Suzuki, A., Ilic, K., Tong, W. (2014). Mining hidden knowledge for drug safety assessment: topic modelling of LiverTox as a case study. BMC Bioinformatics, 15

Zeeman, D., Jones, R., Dysart, J. (2011). Assessing Innovation in Corporate and Government Libraries. Computers in Libraries, 31(5)

Zeng, M.L. (2008). Knowledge Organization Systems (KOS). *Knowledge Organization*, 35(2/3).