The Best of Both Worlds: Highlighting the Synergies of Combining Manual and Automatic Knowledge Organization Methods to Improve Information Search and Discovery

*Paul H. Cleverley and ** Simon Burnett

Robert Gordon University, Department of Information Management, Garthdee Road, Aberdeen AB10 7QB, *<p.h.cleverley@rgu.ac.uk>, **<s.burnett@rgu.ac.uk>

Paul H. Cleverley is a researcher at Robert Gordon University within the Department for Information Management in the Aberdeen Business School. A geoscientist by training and Fellow of the Geological Society, he has worked in the international oil and gas exploration industry for over twenty years. His research focuses on re-examining and re-conceptualizing enterprise search–how we locate and discover data, information and knowledge in the organization.

Simon Burnett is a professor in Robert Gordon University iSchool, and an associate director of the Scottish Graduate School for Social Science. Prior to this, he was Information and Communication Theme Leader in The Research Institute for Management, Governance and Society (IMaGeS). He began his career as a website designer, and established the Centre for Knowledge Management at RGU in 2000. His research has focused on the use of narratives for sharing knowledge, organisational storytelling and the use of social media in professional and socio-political contexts.

Cleverley, Paul H., and Burnett, Simon. **The Best of Both Worlds: Highlighting the Synergies of Combining Manual and Automatic Knowledge Organization Methods to Improve Information Search and Discovery.** *Knowledge Organization. 42(6)*, 428-444. 120 references.

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 was 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.

Received: 30 July 2015; Revised 14 August 2015; Accepted 18 August 2015

Keywords: search, information, manual indexing, automated indexing, KOS, knowledge organization methods

1.0 Introduction

Oil and gas exploration seeks to identify and model hydrocarbon resources through geoscientific methods. Exploration wells can cost over $100 million 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 (McKinsey 2012; Doane 2010; Outsell 2005; IDC 2005; Lowe et al. 2004; Adkins 2003; Delphi 2002; IDC 2001). Much lower figures (9-14%) have been reported from observational studies in organizations (Robinson 2010) and much higher figures (40%) reported (Chum et al. 2011) in the oil and gas industry. A review of surveys indicates that 48% of organizations felt search was unsatisfactory (Norling and Boye 2013; Mindmeter 2011; Doane 2010; Microsoft and Accenture 2010; IDC 2009; AIIM 2008; IDC 2005; Tonstad and Bjorge 2003).

Executives (Oracle 2012) 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. 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 (White 2012; DeLone and McLean 2002) are numerous, including information silos, information literacy, governance and technology issues. 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 (Cleverley and Burnett 2015a) is the minimal use of faceted search and categories, which rarely stimulate serendipitous encounters. 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, 10) describe the decision making process of visionary companies in terms of "the tyranny of the OR, genius of the AND" when coping with contradictory forces. Is this a philosophy to apply to enterprise knowledge organization with respect to manual and automated methods?

Furthermore, knowledge organization systems (KOSs) 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; 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.

2.0 Literature review

This section presents a critical review of the academic and practitioners' literatures relevant to the research. 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.

2.1 Knowledge organization (KO)

Knowledge organization (KO) (Ohly 2012) expresses and imposes a particular structure of knowledge (a "view of reality") behind collections of information. This reality is socially constructed (Berger and Luckman 1966): what is reality for one group may not be for another. Hjørland (2008, 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." Hjørland 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) may contradict the definition made by Hjørland that KO is the preserve of specialists. Corporate library or information center functions (Heye 2003) have traditionally focused on the centralized manual indexing of information using KOS, with indexes under their stewardship. The growth in digital information creation has led to the breakup of these gatekeeping services and the centralized manual indexing model. Zeeman, Jones, and Dysart (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. Classification (Jacob 2004) 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. Applying this to a "typical" oil and gas document, classification may involve assigning an item to a single Document Type, i.e. 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 (Hedden 2013) 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 including those from outside controlled vocabularies to emphasize prominence.

2.2 Knowledge organization systems (KOS)

Hodge (2000, 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 (Alyahyaee 2012) in Oman.

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. Shiri et al. (2002) found evidence of thesauri being both marginal and a substantial source of terms to take advantage of when searching.

From an information search engine perspective (Preece et al. 2001), KOSs can mitigate the vocabulary problem when converted into machine readable forms through knowledge engineering (KE) techniques. 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 (Palmer et al. 2001) can be created automatically through text clustering. 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 challenge 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 (McNaughton 2015, 3), "The pro search, anti-thesaurus motion was defeated resoundingly." 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 KOSs provide benefits. According to Greenberg (2011, 12): "when knowledge structures are absent, the information system is generally considered sub-standard. KOS are a necessity: they inform and promote discovery." Greenberg contrasts this with, "Benefits aside, we must also acknowledge that schemes may reinforce erroneous views, false perceptions and limit new discoveries."

2.3 Information retrieval (IR)

Manning et al. (2009, 1) describe 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 modes 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. These NLP techniques (Manning and Schutze 1999) attempt to understand text in the same way as humans do, catering for the creativity (and ambiguities) in grammar. 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 (Carpineto and Romano 2012) can be achieved linguistically or statistically through creation of an "associative thesauri." 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 (Navigli and Velardi 2002) is its lack of technical domain specific terminology. Examples of corpus dependent techniques (Mikolov et al. 2013, Landauer and Dumais 1997) include the use of term co-occurrences for weighted query expansion and techniques such as latent semantic analysis. These types of search are often termed semantic search.

In a comparison of empirical results, Carpineto and Romano (2012, 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, 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 and ontologies have been used extensively for AQE (Shiri et al. 2002). 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 autocategorize 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 precision, although statistical probabilistic techniques were not applied.

2.4 Enterprise search

Enterprise search (White 2012) 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. Some oil and gas search deployments (Demartini 2007; Behounek and Casey 2007; Palkowsky 2005) index both

structured and unstructured information integrating through meta-models and vocabularies and also enable spatial (map-based) search.

In enterprise search (Andersen 2012), most queries are single word (lookup) and often portrayed as not working well compared to Internet search engines. The crowd using search in enterprises is very small compared to the Internet, hampering the effectiveness of using statistical crowd-sourced usage data for all but the most common of search queries. Many users (Skoglund and Runeson 2009) 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.

Lookup (known item) search (Marchionini 2006) is likely to account for between 80-90% by volume of all enterprise search tasks (Stenmark 2008) 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 (Garbarini et al. 2008, 4): "Any capture of metadata that took more than ten seconds to saving a file was considered problematic." 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?

2.5 Serendipity

Innovation or creativity sparked by an unexpected seemingly random event is often called serendipity. Within organizations, the discovery of innovations and business opportunities (Friedman 2010) is often serendipitous. 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 and Toms 2011) making unforeseen connections. Browsing can support creativity (Bawden 1986), whether the intent is purposive or exploratory in nature. A prerequisite to serendipity (Foster and Ford 2003) is a prepared mind. 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.

2.6 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 (Fagan 2010), displaying an overview of search results, inviting further interaction to filter information and can improve task performance. Low usage of faceted search (5-12%) (Ballard and Blaine 2011; Niu and Hemminger 2010) has also been reported.

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, 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, 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 (Kaizer and Hodge 2005) are typically ordered by the most statistically frequent or most popular. Categories or tags (Cleverley and Burnett 2015b) displayed in faceted search that are representative of an information item (container) as whole, rarely contain intriguing or non-obvious associations.

Information extraction (IE) (Goker and Davies 2009, 132) using search term word co-occurrence is one technique to introduce local context into search refining "what resources are nearby." 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?

2.7 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 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 (Walkup and Ligon 2006) 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. 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.

2.8 Automatic classification/categorization

Classification and categorization (Villena-Roman et al. 2011) can be achieved through machine learning using linguistic rules, labeled training sets or a combination. In a study of legal document categorization, Roitblat et al. (2009, 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 (Meza 2014). Sidahmed et al. applied topic modeling on unstructured text in daily oil and gas drilling reports at BP to identify trending safety issues before they became serious, that were not apparent to engineers. Accuracy was raised as an issue (2015, 10): "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 (Faith 2011) indicating 70% accuracy. 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% (Magnuson 2014) were reported by the US Army, using 11,915 emails as a training set to auto-classify email to 54 records categories. 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 (Hedden 2013, Faith 2011) indicate 50-100 labelled training documents are typically required to give "good" results per category.

Document type classes that require "hard classification" (binary classification to a single class) are probably the most challenging of machine learning document KO tasks; accuracy percentages (Painter et al. 2014) can be as low as 31%. 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 (Wessely 2011) through a manually created linguistic KOS aided by CAKE methods. Topic modeling (Meza 2014) has also been applied to enterprise lessons learnt systems to reveal hidden connections. This (Piantanida et al. 2015) contrasts with a lessons learned system deployed by ENI which focused on a single (manual) categorization approach.

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 applied to the same content 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.

3.0 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, 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 participant comments and nuances.

3.1 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 (Norling and Boye 2013) was chosen because of its size, as some surveys identify search as being more difficult in larger companies. 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.

3.2 Colour coding scheme for methodologies

A colour coding scheme is used in the methodology to describe the interactions between manual and automated methods used in the research (Figure 1).

3.3 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 file name. 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 was performed using a commercially available oil and gas thesaurus licensed by the organization. A geoscience subset of the thesaurus (2,510 concepts) was used for the study including the subject areas of geological time, lithology and depositional environment.

The automated categories were presented in a hierarchical faceted search menu on the left hand side of the search UI to provide an overview of search results, containing a number showing how many documents had been found containing that concept. Search results were listed in the middle of the screen. The value of providing a series of visual "prompts" (facets) to browse and filter search results would be assessed by geoscientists. The methodology is shown in Figure 2.

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

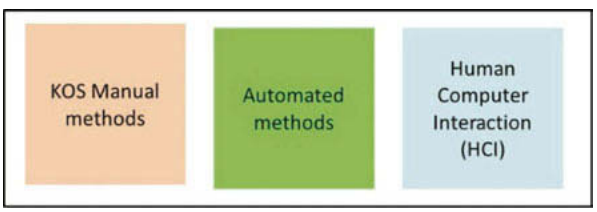

Figure 1–Colour coding scheme used in the study methodology

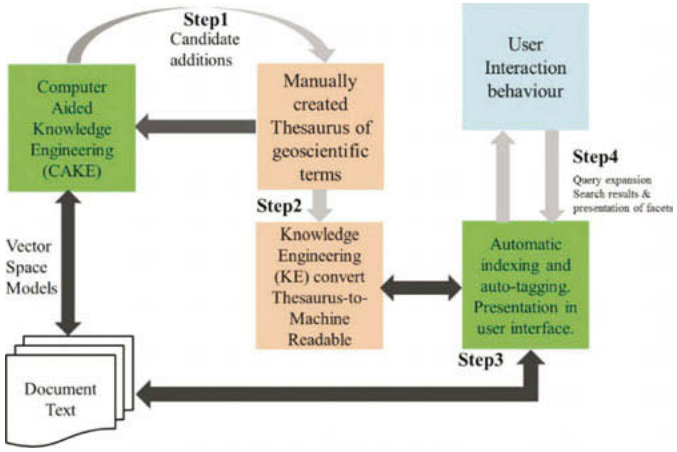

Figure 2–Methodology for research question 1 and 2

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 (Cholakian 2013) used in practice elsewhere. 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 (NSS 2014) was used to evaluate results in order to give 95% confidence.

In Step2 relevant thesaurus associations were made explicit in order to be machine readable, an activity often termed knowledge engineering (KE). This was 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 (e.g. Priabonian OR Bartonian).

The thesaurus with the synonym candidate additions (from Step1) and inference rules (from Step2), was used to automatically 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 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 questions allowing a ranking by Likert scale (1=not at all, 5=to a large extent) and a space for comments:

1. To what extent do these new techniques improve your ability to find information (and why)?
2. What reduction in time spent searching would these techniques make (and why)?
3. To what extent do the techniques allow the discovery of new insights (and why)?
4. Rate the most important features (methods) on display in the user interface (UI) and explain why.

3.4 Research design for question 3

The methodology in Figure 3 address the third research question, to what extent can manual and automated KOS techniques be combined in a search UI to stimulate serendipity.

A semi-interactive "stimulant" was created based on local context. The stimulant was designed to provoke interaction and discussion using content from the Society

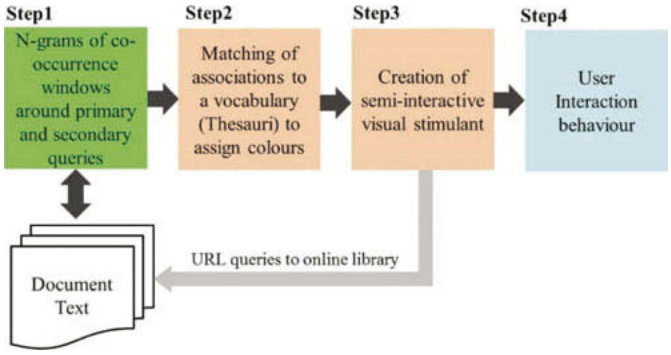

Figure 3–Methodology to address research question 3

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

Figure 4–Word co-occurrences for the primary query "seismic" and secondary query "Malaysia" and "Nigeria" (orange=realm, green=natural process, yellow=matter/materials, blue=property)

of Petroleum Engineers (SPE) in the form of 70,000 article abstracts. Sample search queries were chosen from the study organization's search logs to ensure they were representative.

In Step1 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.

In Step2 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 4).

Step3 created a semi-interactive stimulant. The representation (Halvey and Keane 2007) 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. Fifty associations (rows) (Cleverley and Burnett 2015a) 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 menus.

Algorithm A displayed representative unigrams and Algorithm B bigrams (ranked by descending frequency of occurrence). Algorithm C was a discriminatory set of words 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 to a search query allowing staff to click through and see the search results and documents in which that association occurs.

In Step4 data on user interaction with the stimulant was captured. Sixteen geophysicists took part in the use case study; company staff were purposefully sampled (Coyne 1997) to ensure representation from different geophysical departments. Focus groups (Morgan 1997) were used to enable the researchers to quickly identify a full range of perspectives held by respondents (Powell and Single 1996, 504), "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." The stimulant (Figure 5) was made available on large touchscreens at the organization's premises and participant interactions were video-recorded. Each session lasted 45mins and consisted of between 2-9 staff.

4.0 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.

4.1 To what extent can an oil & gas thesaurus be automatically enhanced (Q1)?

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).

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)

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

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.

4.2 What is the value of auto-categorizing content that is already manually classified (Q2)?

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

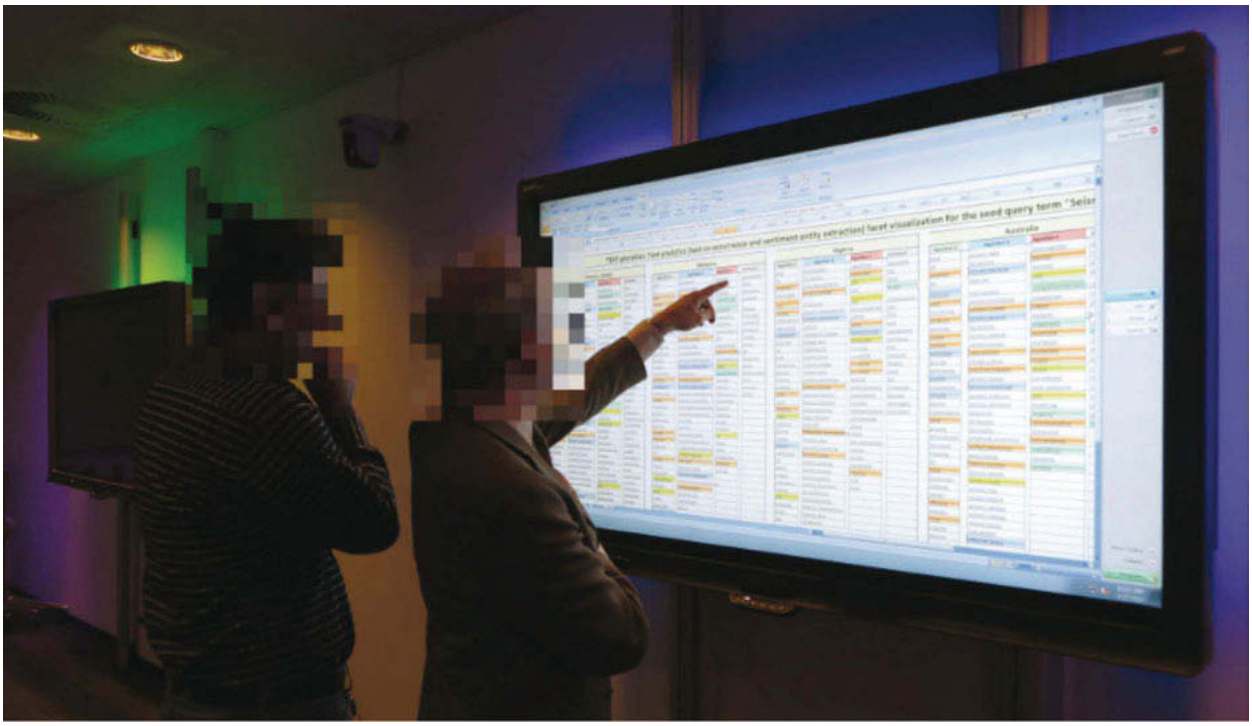

Figure 5–The semi-interactive stimulant on the large touchscreen with participant interactions

(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 menus to support browsing, that manual tagging is unlikely to deliver.

The modal average 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 (Ballard and Blaine 2011, Niu and Hemminger 2010) may contradict research reporting low usage of faceted search. 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, "It would help to find other critical information" [P3].

Some geological basin names found 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); KOSs 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 "play" will not return items on "play" that do not mention. 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 information literacy.

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

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 an 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 is unlikely to deliver.

Personality may influence perceptions of KO techniques, evidenced by "This is overwhelming, too much" [P11] and "Excitement is the first thought" [P28]. All participants preferred Algorithm C, over 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" [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]. A need for search term stimulation was identified."

This helps with big problem (I have with Google), choosing right selection of words to find something." [P13].

5.0 Theoretical Model

A theoretical model (Figure 6) is presented in fulfillment of the research objective.

The KO/KOS methods labelled 1-7 are explained in Table 2, tied to the research questions (Q1, Q2 and Q3). The mixed methods approaches derived from the primary data from this research (methods 3, 5 and 7) are shown in yellow in Figure 6 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) to cater for different content and search goals. The primary data from this research, 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 may predominantly support Lookup search, automated methods supporting predominantly exploratory search.

6.0 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 li-

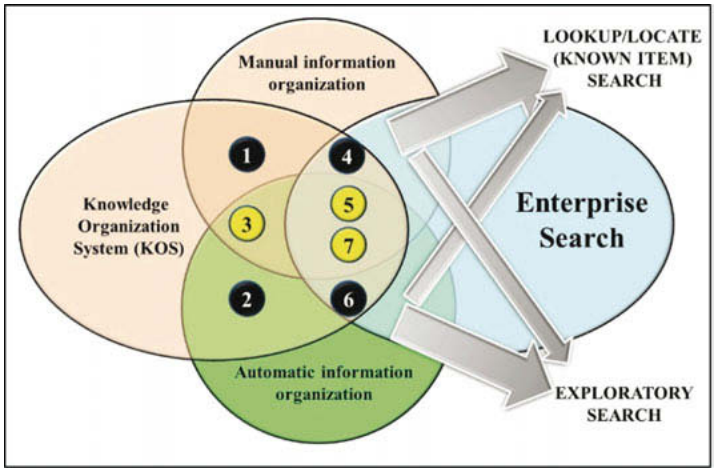

Figure 6–The link between different KO methods and types of enterprise search goal.

	Method	Description
KOS Structure Creation Method	**1. Single method** (Manually developed KOS)	Manual taxonomies and authority lists 'predominantly' support precise lookup search.
	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 clusters and associations from text.
	3. Mixed method (Manual and Automatic): Including results from this study (Q1)	Semi-supervised statistical techniques applied to enterprise content, can enhance the quality of a manually developed KOS.
KO Method	**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 not available in the text).
	5. Mixed method (Automatic classification and/or categorization (to a KOS): Including results from this study (Q2)	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 organized. There is also potential value in applying as a primary method to information that 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 patterns that could be missed if enforcing a KOS.
	7. Mixed method (Manual and Automatic) organization: Including results from this study (Q3)	A manually created KOS can be combined with unsupervised word co-occurrence clustering to facilitate serendipitous information discoveries which a single method may not achieve.

Table 2–KO approaches and link to enterprise search and discovery goals.

brary and information science, data management, KM, IR and AI 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 KOSs. 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 and 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 KOSs 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.

References

Abel, Roger and Paul H. Cleverley 2007. *Improving Information Delivery*. Hart's E&P March Edition. http://www.epmag.com/improve-information-delivery-725091

Adkins, Sam. 2003. *Information Gathering in the Electronic Age: The Hidden Cost of the Hunt*. Safari Techbooks, January 2003.

AIIM 2008. *Findability: The Art and Science of Making Content Easy to Find.* Association for Information and Image Management. http://www.aiim.org/PDFDocuments/34835.pdf.

Allan, James, Bruce Croft, Alistair Moffat and Mark Sanderson. 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, no. 1: 2-32

Alyahyaee, Ali. 2012. "Country report Oil and Gas Data Repository (OGDR) Sultanate of Oman" In *Proceeding of National Data Repository Conference, 21-24 October 2012. Kuala Lumpur, Malaysia.*

Andersen, Espen. 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 of Management* 12, no. 11.

Ballard, Terry and Anna Blaine 2011. "User Search Limiting Behaviour in Online Catalogs. Comparing Classic Catalog Use to Search Behaviour in Next Generation Catalogs." *New Library World* 112, nos. 5/6: 261-73.

Bawden, David. 1986. "Information-Systems and the Stimulation of Creativity." *Journal of Information Science* 12: 203-16.

Behounek, Shawn and Katya Casey 2007. "Earth-Search=GoogleEarth Enterprise+PetroSearch." In *Bytes & Barrels: An Energy Renaissance: Proceeding of 2007 Digital Energy Conference and Exhibition, 11-12th April, Houston, Texas, USA.* Richardson, Texas: Society of Petroleum Engineers.

Berger, Peter and Thomas Luckmann 1966. *The Social Construction of Reality. A Treatise in the Sociology of Knowledge.* 1st ed. London: Penguin.

Bhogal, J., A. Macfarlane and P. Smith. 2007. "A Review of Ontology Based Query Expansion." *Information Processing and Management* 43: 866-86.

Blackman, Sarah. 2012. "Risky Business: Challenges of Deepwater Drilling in the North Sea." *Offshore Technology.com.* http://www.offshore-technology.com/features/featurerisky-business-deepwater-drilling-north-sea/

Caballero, Richard and Steven Nuernberg 2014. "Building an Enterprise Taxonomy." *Presented at the 18th International Petroleum Data, Integration and Data Management, 20-22 May 2014, Houston, TX.*

Carpineto, Claudio and Giovanni Romano 2012. "A Survey of Automatic Query Expansion in Information Retrieval." *ACM Computing Surveys* 44: 1-50.

Cholakian, Andrew. 2013. *How to Use Fuzzy Searches in Elastisearch.* https://www.found.no/foundation/fuzzy-search/

Chuang, Jason, Christopher D. Manning and Jeffrey Heer 2012. "'Without the Clutter of Unimportant Words': Descriptive Keyphrases for Text Visualization." *ACM Transactions on Computer-Human Transactions* 19, no. 3: 1-29.

Chum, Frank, Melanie A. Everett, Scott Hills, Ramakrishna Soma and Roger Cutler 2011. "Realizing the Semantic Web Promise in the Oil & Gas Industry: Challenges and Experiences." *Semantic Technology Conference, 5-9 June 2011, San Francisco, CA.*

Cleverley, Paul 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* 1: 25-32.

Cleverley, Paul H. and Simon Burnett 2015a. "Retrieving Haystacks: A Data Driven Information Needs Model for Faceted Search." *Journal of Information Science* 41: 97-113.

Cleverley, Paul H. and Simon Burnett 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* 14: 1-27.

Collins, Jim C. and Jerry I. Porras 1997. *Built to Last. Successful Habits of Visionary Companies.* New York, HarperCollins.

Coyne, Imelda T. 1997. "Sampling in Qualitative Research. Purposeful and Theoretical Sampling; Merging or Clear Boundaries." *Journal of Advanced Nursing* 26: 623-30.

Dale, Ed. 2013. The Importance of Constant Measurement in Search Relevance. A Longitudinal Case Study. New York: Ernst & Young. Enterprise Search Summit.

DeLone, William H. and Ephraim R. McLean 2002. "The DeLone and McLean Model of Information System Success: A 10 year Update." *Journal of Management Information Systems* 19, no. 4: 9-30.

Delphi. 2002. *Taxonomy & Content Classification. Market Milestone Report.* http://www.delphigroup.com/whitepapers/pdf/wp_2002_taxonomy.pdf

Demartini, Gianluca. 2007. "Leveraging Semantic Technologies for Enterprise Search." In *PIKM '07 Proceedings of the ACM first Ph.D. workshop in CIKM 5-10 November 2007, Lisboa, Portugal.* New York: ACM, 25-32.

Doane, Mike. 2010. "Cost Benefit Analysis: Integrating an Enterprise Taxonomy into a Sharepoint Environment." *Journal of Digital Asset Management* 6: 262-78.

Fagan, Jody C. 2010. "Usability Studies of Faceted Browsing: A Literature Review." *Information Technology and Libraries* 29: 58-66.

Faith, Ashleigh. 2011. "Linguistically Training Automatic Indexing Software for Complex Taxonomies." *Presented at the Semantic Technology & Business Conference 2-5 June 2013, San Francisco, CA.* https://ashleighnfaith.files.wordpress.com/2014/01/semtech2013_a-faith_pdf.pdf

Foster, Allen and Nigel Ford 2003. "Serendipity and Information Seeking: An Empirical Study." *Journal of Documentation* 59: 321-40.

Friedman, Barry. 2010. "Serendipity is an Explorationists Best Friend." *American Association of Petroleum Geologists.* https://www2.aapg.org/explorer/2010/04apr/mobilebay0410.cfm

Furnas, G. W., T. K. Landauer, L. M. Gomez and S. T. Dumais. 1987. "The Vocabulary Problem in Human-System Communication." *Communications of the ACM* 30, no. 11: 964-71.

Garbarini, Mike, Robert E. Catron and Bob Pugh 2008. "Improvements in the Management of Structured and Unstructured Data." *Society of Petroleum Engineers*, Report IPTC12035.

Goker, Ayse and John Davies 2009. *Information Retrieval: Searching in the 21st Century.* UK: Wiley & Sons Ltd.

Greenberg, Jane. 2011. "Introduction: Knowledge Organization Innovation: Design and Frameworks." *Bulletin American Society for Information Science and Technology* 37, no. 4: 12-4.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Generation.* MA: Kluwer Academic Publishers Norwell.

Gwizdka, Jacek. 2009. "What a Difference a Tag Cloud Makes: Effects of Tasks and Cognitive Abilities on Search Results Interface Use." *Information Research* 14, no. 4.

Halvey, Martin J. and Mark T. Keane 2007. "An Assessment of Tag Presentation Techniques." In *Proceedings of 16th International World Wide Web Conference, 8-12 May 2007, Banff, AB, Canada.* New York: ACM, 1313-14.

Hearst, Marti A. and Emilia Stoica 2009. "NLP Support for Faceted Search Navigation in Scholarly Collections." In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, 7th August 2009, ACL-IJCNLP Suntec, Singapore.* Stroudsburg, PA: Association for Computational Linguistics, 62-70.

Hedden, Heather. 2013. "Taxonomies for Auto-Tagging Unstructured Content." *Presented at Text Analytics World, September 30th - October 1st 2013, Boston, MA.*

Heye, Dennie. 2003. "Taxonomies and Automatic Classification at Shell – A Case Study." *Presented at the Building a Knowledge Framework: Practical Taxonomy Design and Application Conference, 29-30 September 2003, Amsterdam, The Netherlands.*

Hjorland, Birger. 2008. "What is Knowledge Organization (KO)?" *Knowledge Organization* 35: 86-101.

Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files.* Washington, DC: Digital Library Federation, Council on Library and Information Resources.

Hubert, Cindy. 2012. "Seamless Collaboration. Enabling Employees to Work Together Across Boundaries." *APQC Report K03906*, 1-15.

IDC 2001. "The High Cost of Not Finding Information." http://www.ejitime.com/materials/IDC%20on

%20The%20High%20Cost%20Of%20Not%20Finding%20Information.pdf.

IDC 2005. "The Hidden Costs of Information Work. International Data Corporation (IDC)." http://www.slideshare.net/PingElizabeth/the-hidden-costs-of-information-work-2005-idc-report.

IDC 2009. "IDC Executive Briefings: Information Advantage: Information Access in Tommorow's Enterprise. International Data Corporation (IDC)." http://www.3ds.com/fileadmin/PRODUCTS/SIMULIA/e-Booth/PDF/Exalead_IDC-Information_Access_in_Tomorrows _Enterprise-AR.pdf.

Jacob, Elin K. 2004. "Classification and categorization: A Difference that Makes a Difference." *Library Trends* 52, no. 3: 515-40.

Jacobs, Paul S. and Lisa R. Rau 1990. "SCISOR: Extracting Information from On-Line News." *Communications of the ACM* 33, no. 11: 88-97.

Jurka, Timothy P., Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt 2013. "RTextTools: A Supervisory Learning Package for Text Classification." *The R Journal* 5, no. 1: 6-12.

Kaizer, Jasper and Anthony Hodge 2005. "AquaBrowser Library: Search, Discover, Refine." *Library Hi Tech News* 22, no. 10: 9-12.

Krestel, Ralf, Gianluca Demartini and Eelco Herder 2011. "Visual Interfaces for Stimulating Exploratory Search." In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, 13-17 June 2011, Ottawa, Canada.* New York: ACM, 393-4.

Landauer, Thomas K. and Susan T. Dumais 1997. "A Solution to Platos' Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review* 104: 211-40.

Lennon, Andrew, Fatima Alshubi and Paul H. Cleverley 2012. "Improving Subsurface and Wells Document Management at Qatar Shell." *Presented at 16th Annual Petroleum Data Integration Conference. 15-17 May 2012, Houston, USA.*

Low, Boon. 2011. *Usability and Contemporary User Experiences in Digital Libraries.* CIGS Seminar, University of Edinburgh.

Lowe, Alistair, Chris McMahon and Steve Culley. 2004. "Characterising the Requirements of Engineering Information Systems." *International Journal of Information Management* 24: 401-22.

Lykke, Marianne and Anna G. Eslau 2010. "Using Thesauri in Enterprise Settings: Indexing or Query Expansion?" In *The Janus Faced Scholar: A Festschrift in Honour of Peter Ingwersen*, edited by Peter Ingwersen, Birger Larsen, Fredrik Åström and Jesper W. Schneider, 87-97. Det Informationsvidenskabelige Akademi.

Magnuson, Doug. 2014. "Auto Classification and the Holy Grail for Records Managers." *IBM Presentation as the Association or Records Managers and Administrators (ARMA), Houston.* http://c.ymcdn.com/sites/www.armahouston.org/resource/resmgr/conference/session_tu3b_auto_classifica.pdf

Manning, Christopher D. and Hinrich Schutze 1999. *Foundations of Statistical Natural Language Processing.* Cambridge: Massachusetts Institute of Technology Press.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schutze 2009. *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Marchionini, Gary. 2006. "Exploratory Search: From Finding to Understanding." *Communications of the ACM* 49, no. 4: 41-6.

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

McCandless, David. 2012. *Information in Beautiful*, 2nd ed., William Collins, London.

McCay-Peet, Lori and Elaine Toms 2011. "Measuring the Dimensions of Serendipity in Digital Environments." *Information Research* 16, no. 3.

McKinsey 2012. *The Social Economy: Unlocking Value and Productivity through Social Technologies.* McKinsey Global Institute Report. http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_social_economy

McNaughton, Neil. 2015. "Knowledge Organization – The Great Debate!" *Oil Information Technology Journal* 20, no. 2: 1-11.

Meza, David. 2014. *On Developing Better Magnets for Needles in Haystacks.* Office of the Chief Knowledge Officer (CKO), National Aeronautical Space Administration (NASA). http://km.nasa.gov/on-developing-better-magnets-for-finding-needles-in-haystacks/

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

Mikolov, Tomas, Ilya Sutskever, Kai Chen, K., Greg. S. Corrado and Jeffrey Dean 2013. "Distributed Representations of Words and Phrases and their Compositionality." *Advanced in Neural Information Processing Systems* 26: 3111-9.

Miller, Don. 2014. *Just the facts Auto-classification and Taxonomies* http://www.slideshare.net/martingarland1/just-the-facts-autoclassification-and-taxonomies-webinar

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine Miller. 1990. "Wordnet: An Online Lexical Database." *International Journal of Lexicography* 3: 235–44.

Mindmeter 2011. "Mind the Enterprise Search Gap." *Report Sponsored by SmartLogic.*

Morgan, David L. 1997. "Planning and Research Design for Focus Groups." In *Focus Groups as Qualitative Research, 2nd ed.*, 32-46. Thousand Oaks, Calif.: Sage.

Morville, Peter and Louis Rosenfeld 2006. *Information Architecture for the World Wide Web: Designing Large-Scale Websites*. 3rd ed. Beijing: O'Reilly.

Munkvold, Bjorn E., Terro Paivarinta, Anne K. Hodne and Elin Stangeland 2006. "Contemporary Issues of Enterprise Content Management: The Case of Statoil." *Scandinavian Journal of Information Systems* 18: 69-100.

Navigli, Roberto and Paola Velardi 2002. "Automatic Adaptation of WordNet to Domains." In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC '02), 29-31 May 2002, Canary Islands, Spain.*

Niu, Xi and Bradley M. Hemminger 2010. "Beyond Text Querying and Ranking List: How People are Searching through Faceted Catalogs in Two Library Environments." In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem* 47, no. 1: 1-9.

Noor, Azahar M. and Che Zan H. Yassin 2006. "Issues, Challenges and Constraints in K-Era." In *Proceedings of the Knowledge Management International Conference, 6-8th June 2006, Kuala Lumpur, Malaysia.*

Norling, Kristian and J Boye 2013. "2013 Findability Survey." *Findability Day, May 30 2013, Findwise, Stokholm.*

NSS 2014. "National Statistics Service Australia Online Calculator." *National Statistical Service.*

O'Donnell, Mick. 2011. "Visualizing Patterns in Text." *Keynote talk at AESLA (Spanish Association of Applied Linguistics), 4-6 May 2011, University of Salamanca.*

Ohly, H. Peter. 2012. "Organization, Management and Engineering of Knowledge: Rivals or Complements?" In *20 years of ISKO Spanish Chapter: Proceeding of X Congress ISKO Spanish Chapter, June 30 –July 1 2011,* edited by María del Carmen Pérez Pais and María G Bonome. Ferrol: Universidade da Coruña, Servizo de Publicacións, 541-51.

Oil and Gas UK 2011. "Oil and Gas UK." *Exploration Economic Report 2011.* http://www.oilandgasuk.co.uk/cmsfiles/modules/publications/pdfs/EC027.pdf.

Oracle 2012. "From Overload to Impact: An Industry Scorecard on Big Data Business Challenges." http://www.oracle.com/us/industries/oracle-industries-scorecard-1692968.pdf.

Outsell 2005. "Survey of Knowledge Workers." http://www.outsellinc.com.

Painter, Kyle, Steven J. Dutton, Elizabeth O. Owens and Lyle D. Burgoon 2014. Automatic Document Classification for Environmental Risk Assessment. *PeerJ* PrePrints 2:e300v1. https://dx.doi.org/10.7287/peerj.preprints.300v1.

Palkowsky, Betsy. 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., J. Pesenti, R. E. Valdes-Perez, M. G. Christel, A. G. Hauptmann, D. Ng, H. D. Wactlar. 2001. "Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results." In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, 24-28 June 2011, Roanoke, VA.* New York: ACM, 451.

Peng, Jie, Ben He and Iadh Ounis 2009. "Predicting the Usefulness of Collection Enrichment for Enterprise Search." In *Advances in Information Retrieval Theory,* edited by L. Azzopardi et al., 366-70. Berlin: Springer-Verlag.

Piantanida, Marco, Elena Cheli, Onest Gheorghisor, Paolo Rossi 2015. "Processes and Tools to Effectively Leverage on Lessons Learned for E&P Development Projects." *Presented at Offshore Mediterranean Conference and Exhibition, 25-27th March 2015, Ravenna, Italy.*

Powell, Richard A. and Helen M. Single 1996. "Methodology Matters – V." *International Journal for Quality in Health Care* 5: 499-504.

Preece, Alun, Alan Flett, Derek Sleeman, David Curry, Nigel Meany and Phil Perry. 2001. "Better Knowledge Management through Knowledge Engineering." *IEEE Intelligent Systems* 16: 36-42

Quaadgras, Anne and Cynthia M. Beath 2011. "Leveraging Unstructured Data to Capture Business Value." *Center for Information Systems Research. MIT, Sloan School of Management* 11, no. 4.

Raskin, Rob. 2011. *National Aeronautical Space Administration (NASA) Semantic Web for Earth and Environmental Terminology (SWEET) Ontology*. https://sweet.jpl.nasa.gov/

Rasmus, Daniel W. 2013. "How IT Professionals can Embrace the Serendipity Economy." *Harvard Business Review.* https://hbr.org/2013/08/how-it-professionals-can-embrace-the-serendipity/

Robinson, Mark A. 2010. "An Empirical Analysis of Engineer's Information Behaviors." *Journal of the American Society for Information Science and Technology* 61: 640-58.

Roitblat, Herbert L., Anne Kershaw and Patrick Oot 2009. "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review." *Journal of the Association for Information Science and Technology* 61: 70-80.

Romero, Lee. 2013. "Deloitte: Improving Findability in the Enterprise." *Presented at APQC Knowledge Management Conference May 3rd 2013, Houston, Texas, USA.*

Rose, Donna G. 2010. "Apache Corporation's the ECM Journey." *Association for Information and Image Management.* http://www.aiim.org/Documents/chapters/southwest/Donna%20Rose_AIIM_SW-Chapter_05-06-2010.pdf.

Salmador-Sanchez, Maria P. and Maria Angeles Palacios. 2008. "Knowledge-based Manufacturing Enterprises: Evidence from a Case Study." *Journal of Manufacturing Technology Management* 19: 447-68.

Salthe, Stanley N. 2012. "Hierarchical Structures." *Axiomathes* 22: 355-83.

Sasaki, Yutaka. 2008. *Automatic Text Classification*. University of Manchester. http://www.nactem.ac.uk/dtc/DTC-Sasaki.pdf

Schlumberger. 2008. *Schlumberger Oilfield Glossary*. http://www.glossary.oilfield.slb.com/

Shiri, A. A., C. W. Revie, G. Chowdhury. 2002. "Thesaurus-Assisted Search Term Selection and Query Expansion: A Review Of User-Centred Studies." *Knowledge Organization* 29: 1-19.

Sidahmed, Mohamed, Christopher J. Coley and Shawn Shirzadi. 2015. "Augmenting Operations Monitoring by Mining Unstructured Drilling Reports." *Society of Petroleum Engineers SPE-173429-MS*.

Skoglund, Mats and Per Runeson. 2009. "Reference-based Search Strategies in Systematic Reviews." In *Proceedings of the 13th International Conference on Evaluation and Assessment in Software Engineering (EASE). 20-21st April 2009, Durham University*. UK: British Computer Society Swinton, 31-40.

Smiraglia, Richard P. and Charles van den Heuvel 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* 37: 43-7.

Smith, Reid. 2012. "Implementing Enterprise Information Management at Marathon Oil." *Presented at Gartner Portals, Content and Collaboration Summit. Track B: Content and Information Management Session B2, 12th March 2012.* http://www.reidgsmith.com/Case_Study_-_Implementing_Enterprise_Content_Management_at_Marathon_Oil.pdf.

Solskinnsbakk, Geir and Jon A. Gulla 2008. "Ontological Profiles as Semantic Domain Representations." In *Natural Language and Information Systems: 13th International Conference on Applications of Natural Language to Information Systems, 24-27 June 2008, London, UK*, edited by Epaminondas Kapetanios; Vijayan Sugumaran; Myra Spiliopoulou, Berlin; New York : Springer 67-78.

Stenmark, Dick. 2008. "Identifying Clusters of User Behaviour in Intranet Search Engine Log Files." *Journal of the American Society for Information Science and Technology* 59: 2232-43.

Strauss, Anselm and Juliet M. Corbin 1998. *Basics of Qualitative Research. Techniques and Procedures for Developing Grounded Theory.* 2nd ed. Thousand Oaks: Sage.

Tonstad, Kjetil and Eldar Bjorge 2003. "Data Management Metrics in Statoil, Smi Data Management Presentation." London, UK.

Velardi, Paola, Roberto Navigli, Stefano Martinez and Juana Ruiz Martinez 2012. "A New Method for Evaluating Automatically Learned Terminological Taxonomies." In *Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC 2012), May 21-27th, 2012, Istanbul, Turkey,* edited by Nicoletta Calzolari et al.

Villena-Roman, Julio, Sonia Collada-Perez, Sara Lana-Serrano abd Jose Gonzalez-Cristobal 2011. "Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization." In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 8-20 May 2011, Palm Beach, Florida*, edited by R. Charles Murray and Philip M. McCarthy. Menlo Park, Calif.: AAAI Press, 323-8.

Walkup, Gardner W. and Bob J. Ligon 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.

Wessely, Jim. 2011. *Text Analytics and Auto-Categorization in Semantic Web Applications.* http://semtech2011.semanticweb.com/uploads/handouts/Wessely%20SemTech%202011%20color_3988_1996.pdf

White, Martin. 2012. *Enterprise Search.* 1st Edition. California: O'Reilly.

White, Martin. 2014. *Search Strategy A-Z List of Topics. Intranet Focus.* http://www.intranetfocus.com/wp-content/uploads/Intranet-Focus-Search-Strategy-A-Z.pdf

Zeeman, Deane, Rebecca Jones and Jane Dysart. 2011. "Assessing Innovation in Corporate and Government Libraries." *Computers in Libraries* 31, no. 5.

Zeng, Marcia L. 2008. "Knowledge Organization Systems (KOS)." *Knowledge Organization* 35: 160-82.