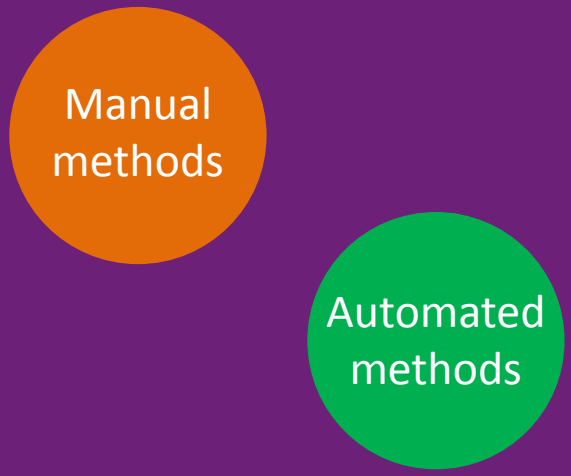
ROBERT GORDON UNIVERSITY ABERDEEN

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 and Simon Burnett

Department of Information Management
Aberdeen Business School
Robert Gordon University (RGU), United Kingdom



Agenda

- Background
- Development of 3 Research questions
- Input to Theoretical Model
- Methodology (Results follow each question)
- Theoretical Model
- Conclusions

Search Analogy: Looking for fossils

BACKGROUND – MOTIVATION FOR RESEARCH

Looking for what I know

Ammonite

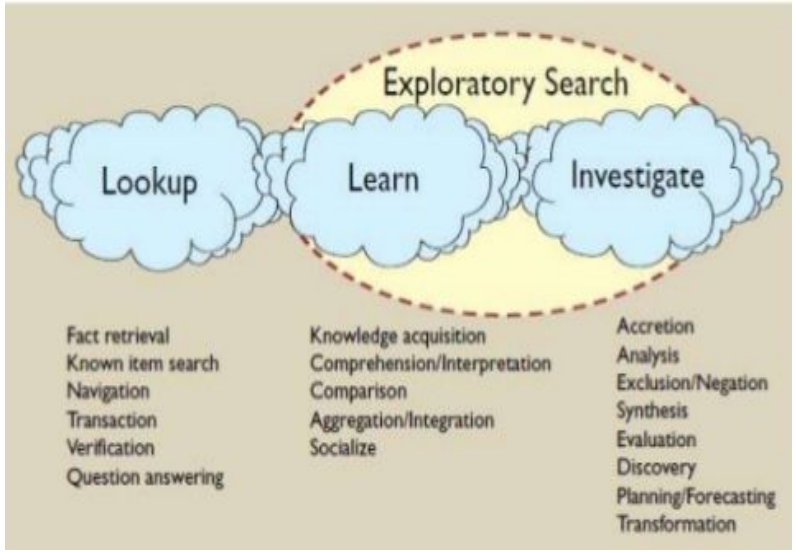
Ichthyosaur
Paddle Bone

Belemnite: https://ferrebeekeeper.wordpress.com/tag/belemnite/
Ammonite: http://www.psychiccowgirl.com/ammolite-albertas-gemstone/

Unexpected - what I don't know

BACKGROUND – MOTIVATION FOR RESEARCH

In an enterprise setting



Marchionini 2006

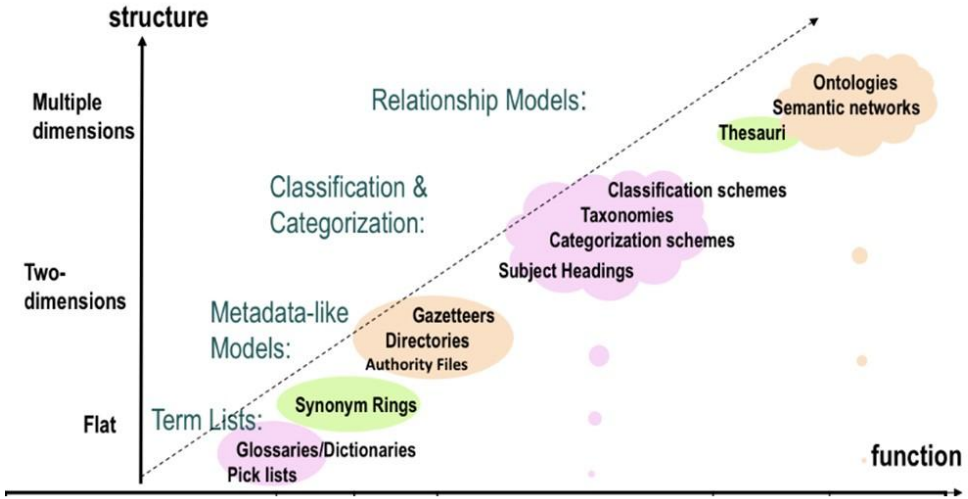
Link between KO/KOS literature and search goals (business requirements for search)?

Enterprise Search & Discovery issues

- **24%** of a professionals time spent looking for information, **48%** of organizations feel search is unsatisfactory in some way.
- Recent research on exploratory search indicates even the most experienced searchers can miss 73% of high value items.
- Executives indicate missed opportunities by failing to leverage their information effectively could represent 22% of annual revenue.

Knowledge Organization (KO)

"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" (Horland 2008)



Taxonomy of Knowledge Organization Systems (KOS) (Zeng 2008)

BACKGROUND – KO/KOS supporting search and discovery?

- The role of thesauri in modern day IR being questioned (ISKO 2015)
- Internet/Enterprise search differences not always recognized (White 2012)
- Traditional corporate libraries have been downsized (Zeeman et al 2011)
- IT departments and software vendors heavily promoting auto-classification and auto-categorization techniques (automation) but not necessarily taking a holistic view
- KOS may promote new discoveries, but may limit others (Greenberg 2011)

DEVELOPMENT RESEARCH QUESTION 1

Taxonomies remain crucial to the oil and gas industry to enable browsing & to support search accuracy

Organizations sometimes treat KO methods (manual/automatic) as mutually exclusive. "*Tyranny of OR*" (Collins and Porras 1997)

Automatic thesaurus construction

- Automated thesaurus creation and enrichment techniques from text corpora are well documented (Grefenstette 1994) although little research applied in the oil and gas industry.

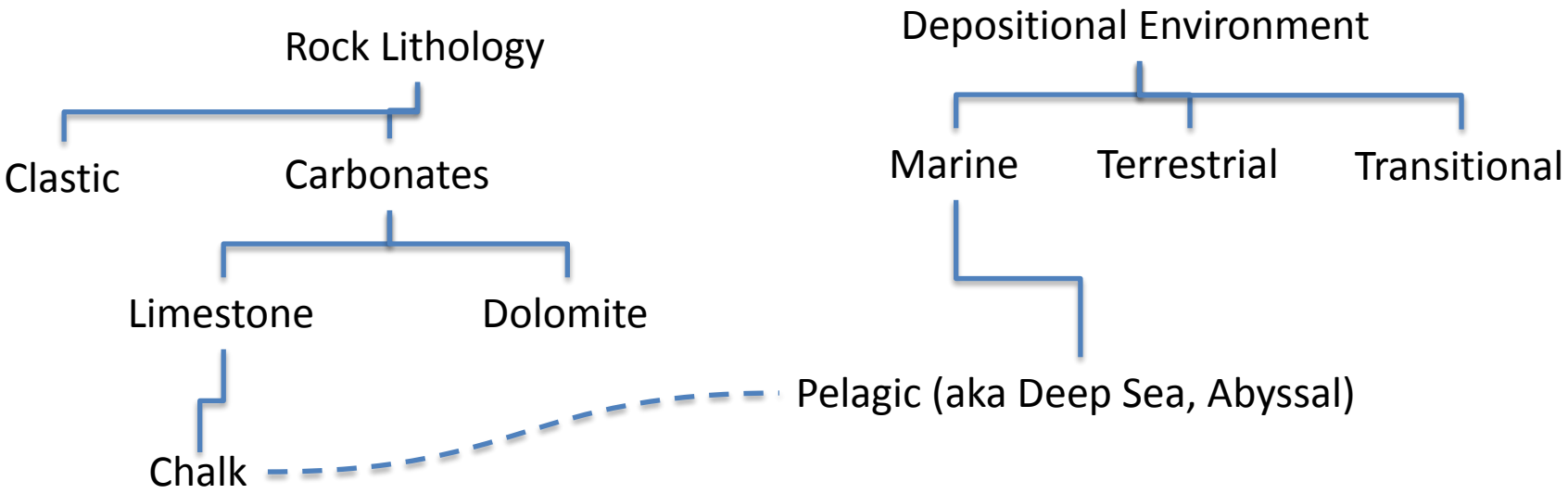
- Velardi *et al.* (2012) stated it is virtually impossible to recreate complex domain specific taxonomies automatically from document content alone.

Research questions

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

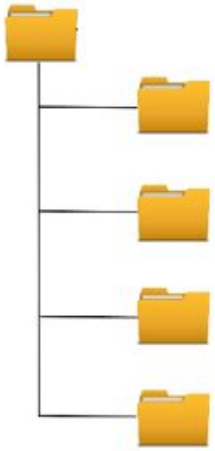
DEVELOPMENT RESEARCH QUESTION 2

Semantic and vocabulary problem when searching: Automatic Query Expansion (AQE)



DEVELOPMENT RESEARCH QUESTION 2

Organizing in Folders and/or Tagging inside EDMS systems (e.g. SharePoint)



"Any capture of metadata that took more than ten seconds to saving a file was considered problematic" (Exxonmobil Garbarini *et al.* 2008)

Finding information just browsing folders can hamper discoverability of certain information. In an EDMS, many end users may not add many tags (if any) affecting search.

Research questions

- 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?**

Serendipity

- Serendipity – Fortuitous information encountering
- Favours prepared mind (Foster and Ford 2003)
- Information rich environments (McCay-Peet and Toms 2011)
- Unlikely to be controllable but developing a capability that may lead to more serendipitous encounters is deemed plausible

DEVELOPMENT RESEARCH QUESTION 3

Browsing can support creativity (Bawden 1986) and lead to serendipitous encounters

(Yang and Wagner 2014)

Search term word co-occurrence filters *may* (Gwizdka 2009, Olsen 2007) or *may not* (Low 2011) aid information discovery.

The unexpected

Most statistically frequent or most popular associations as search filters "*relevant but not interesting*" (Cleverley and Burnett 2015)

Research questions

- 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?**

Additional from literature

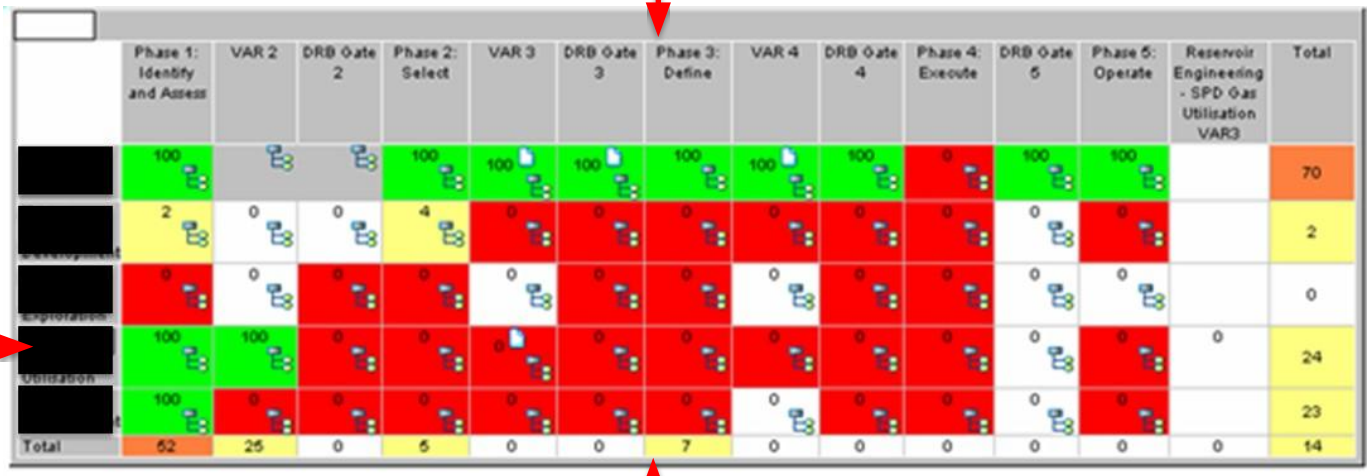
- Manual (pre-attribution)
- Automated Semi-supervised (Linguistic)
- Automated Semi-supervised (Statistical)
- Automated Unsupervised

Stage gate deliverables – Shell

Accurate classification and re-use

Stage gate process (Execution, Assurance, Decisions)

Opportunities
Prospects or
Projects



	Phase 1: Identify and Assess	VAR 2	DRB Gate 2	Phase 2: Select	VAR 3	DRB Gate 3	Phase 3: Define	VAR 4	DRB Gate 4	Phase 4: Execute	DRB Gate 5	Phase 5: Operate	Reservoir Engineering - SPD Gas Utilisation VAR3	Total
[redacted]	100			100	100	100	100	100	100	0	100	100		70
[redacted]	2	0	0	4	0	0	0	0	0	0	0	0		2
[redacted]	0	0	0	0	0	0	0	0	0	0	0	0		0
[redacted]	100	100	0	0	0	0	0	0	0	0	0	0	0	24
[redacted]	100	0	0	0	0	0	0	0	0	0	0	0		23
Total	52	25	0	5	0	0	7	0	0	0	0	0	0	14

Process status and Deliverables (drag and drop)

(Abel and Cleverley 2007)

INPUT TO THEORETICAL MODEL: AUTOMATIC (LINGUISTIC)

Auto-categorize discussions, best practices.

(Wessely 2011)

Weighted (Scored) linguistic rules

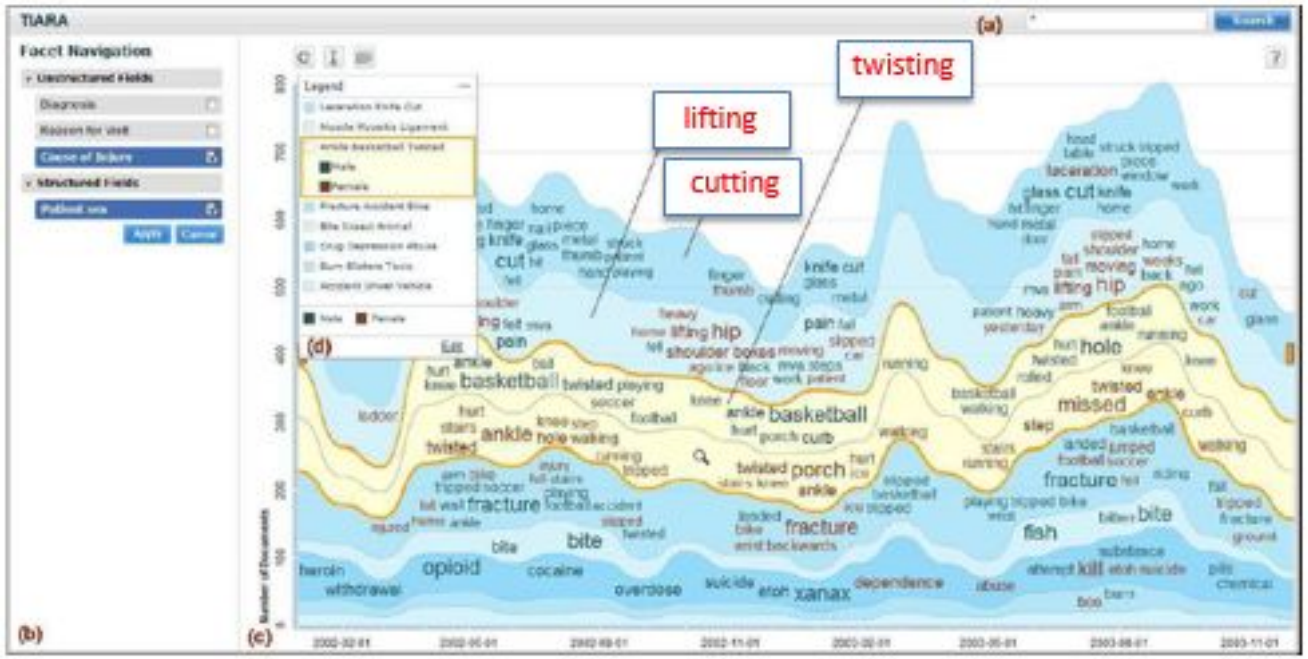
Drilling problem		Score
-Stuck pipe	50	
-Lost circulation		50
-Mud losses	40	
-Fishing	30	
-Gas kick	30	
-Lost time incident		25
-Cost overrun		25
-Schedule delay		25
-Drilling	25	

Using labelled training data

- For rapid, diverse and high volumes of information, manual efforts **costly**
- Reuter's newswires **9,603** training docs to 11 categories (Sasaki 2008)
- US Army **11,915** emails as a training set auto-classify email to 54 records categories, **60-90%** accuracy (Magnuson 2014).
- Practitioner heuristics indicate **50-100** labelled training docs typically required to give good results per category (Hedden 2013, Faith 2011).
- *Hard classification* can be as low as **31%** (Painter *et al.* 2014)
- *Best results from hybrid [linguistic & statistical] methods (*Carpineto & Romano 2012)

INPUT TO THEORETICAL MODEL: AUTOMATIC (UNSUPERVISED)

Topic Modelling (very complex text co-occurrence)

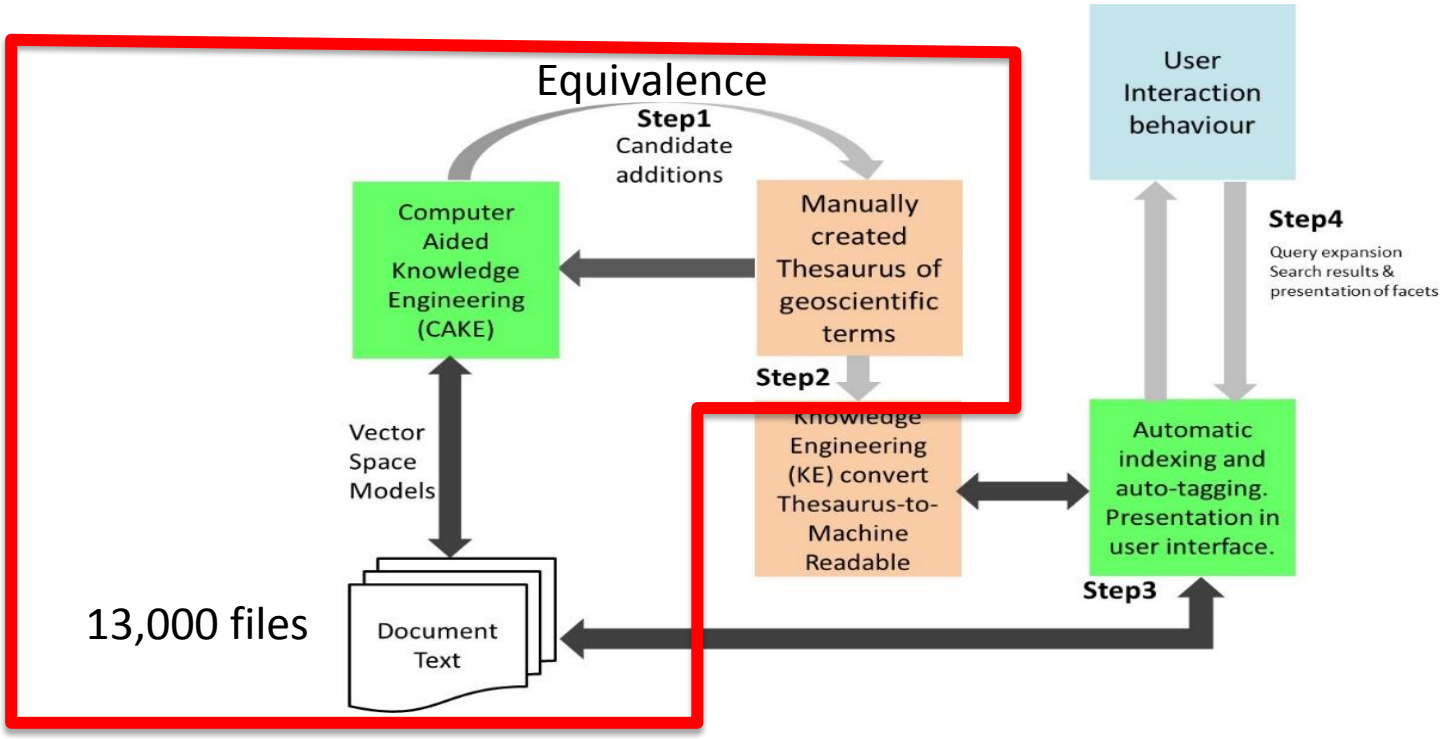


23,000 emergency room records (Wei *et al.* 2010)
Text in '*reason for visit*', '*cause of injury*', '*diagnosis*' fields

METHODOLOGY, SAMPLING and ANALYSIS

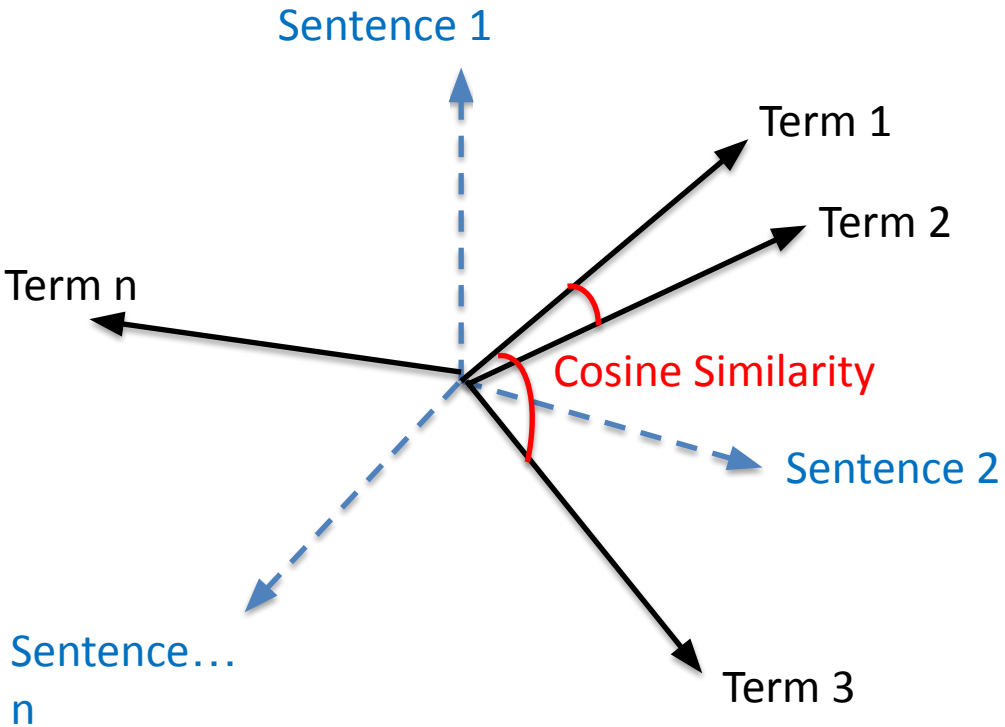
- Pragmatic approach
- Case study oil and gas industry: Representative organization
- Question 2 (6 geoscientists volunteered)
- Question 3 (16 geoscientists purposefully sampled)
- Due to small sample size (caused by organizational changes), subsequent face to face sessions with an additional 12 Geoscientists provided further information for Q2 and Q3
- Analysis is therefore mainly qualitative (Thematic mapping).

Methodology Question 1



METHODOLOGY

Vector space applied to Geoscience text



(After Salton 1975)

Identifying synonyms, lexemes and spelling mistakes through statistics

Seed is Existing Thesaurus licensed by organization
2,500+ concepts

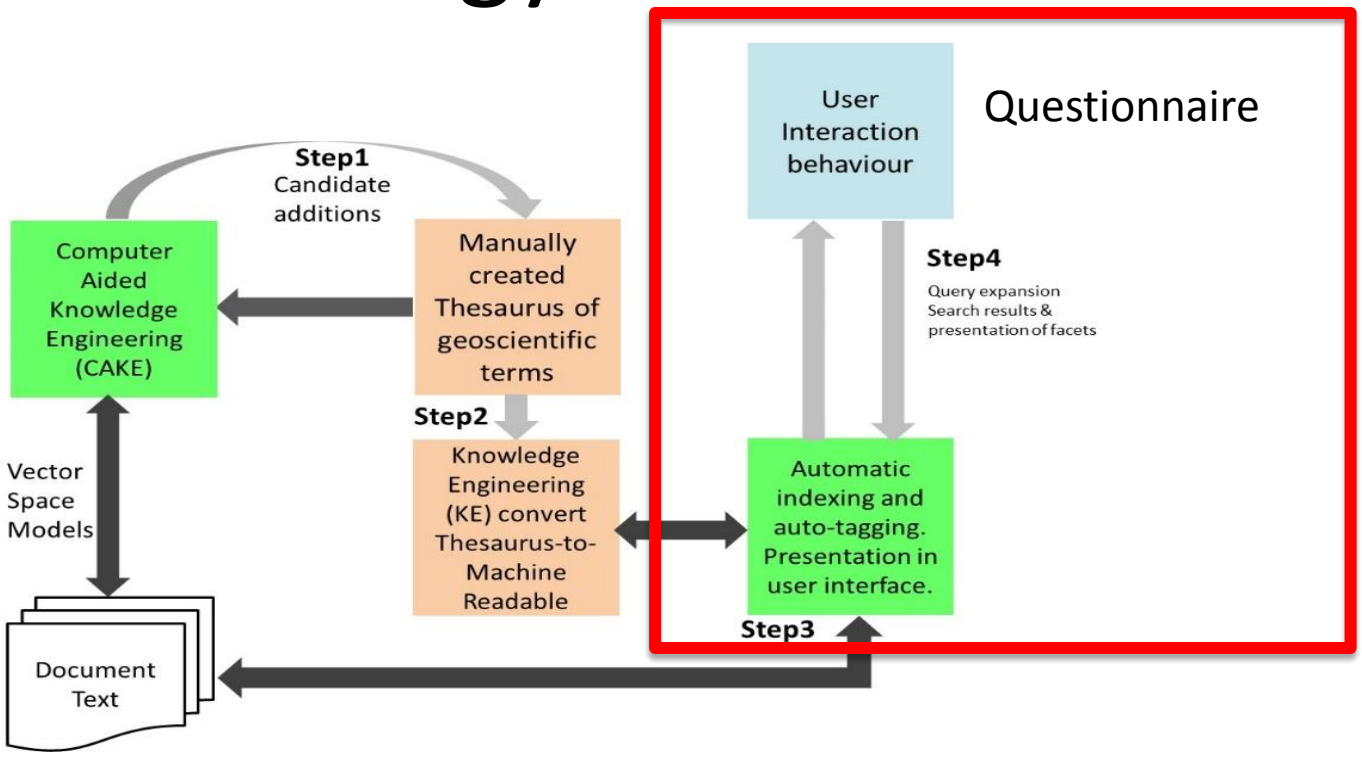
+ Igneous
  - Intrusive Igneous
    - Monzonite

RESULTS – QUESTION 1

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

Sampled 334 concepts from 2,520 to get a 95% confidence figure of a 34% increase in valid lexemes, new synonyms and variants

Methodology Question 2

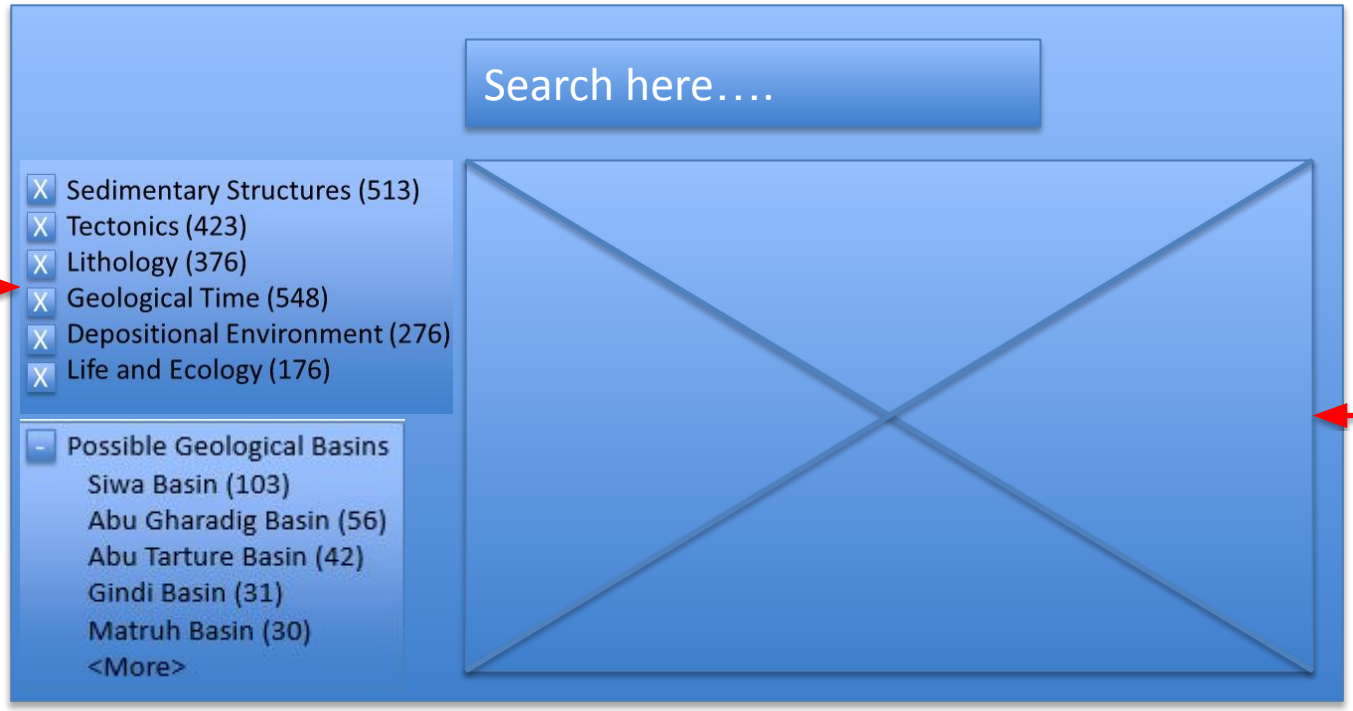


13,000 files

METHODOLOGY

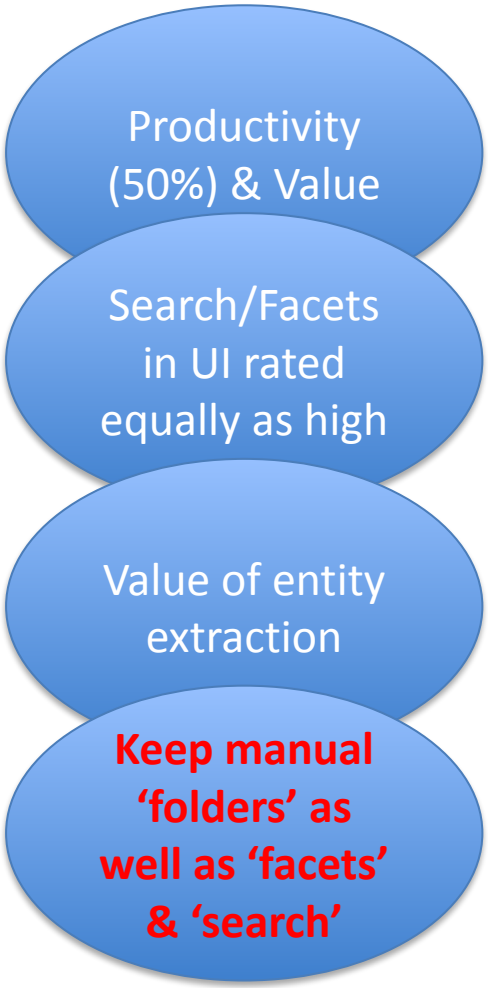
Enterprise Search User Interface

Automatic Hierarchical Faceted Search Refiners



Search Results List

RESULTS – QUESTION 2



"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]

RESULTS – QUESTION 2

- Average number of unique tags added per document by auto-categorization (leaf only)=**113.9** (PDF), **23.25** (Other office files)
- Average number of tags added manually by geoscientists in SharePoint (**3.6** for 2 mandatory pick lists, **1.1** for optional)
- **Value of auto-categorization to increase richness of tags for faceted search to enable browsing and discovery**

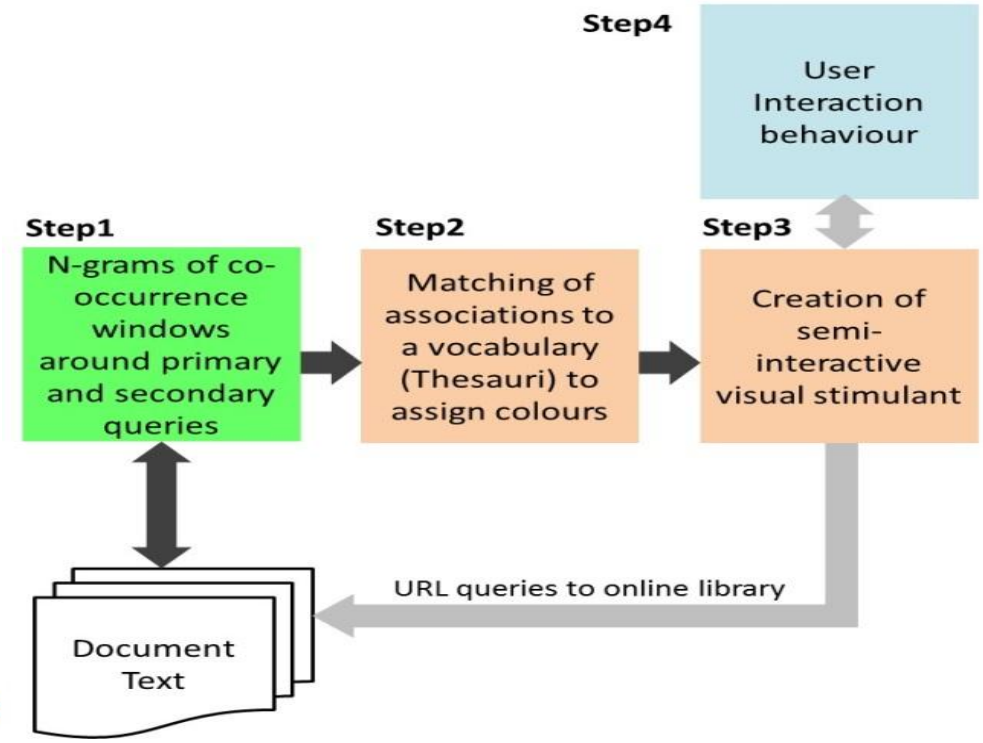
RESULTS – QUESTION 2

"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]

"I learnt that Google is not a Geologist" [P23]

Methodology – Question 3



70,000 papers

Society of Petroleum Engineers

METHODOLOGY

Part of stimulant

PRIMARY

Seismic

SECONDARY

Malaysia

Nigeria

Australia

Canada

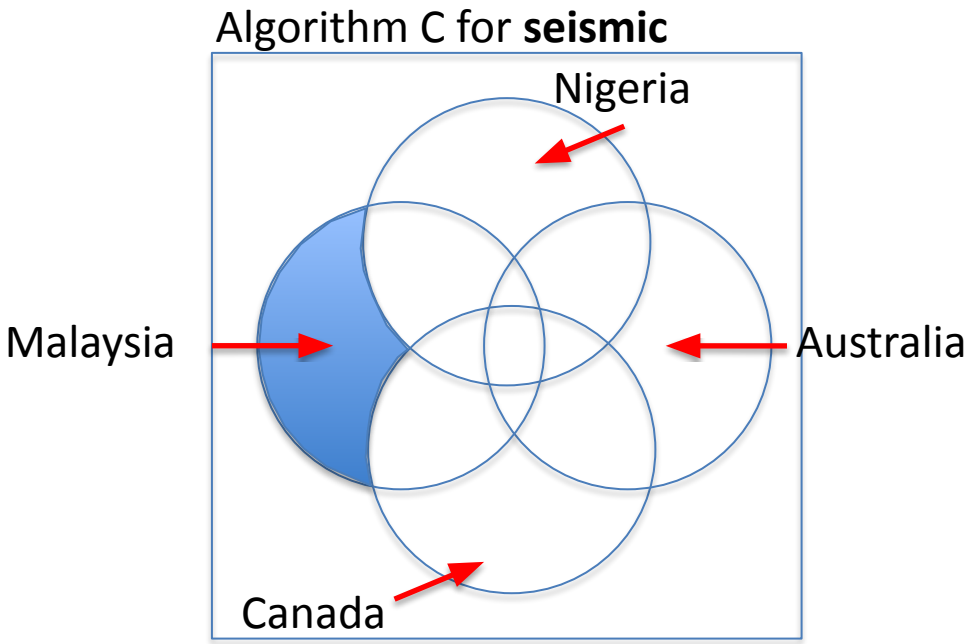
For the primary search query='seismic'

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

METHODOLOGY

Primary search query=seismic,
Secondary Queries=(Malaysia, Nigeria, Australia, Canada)

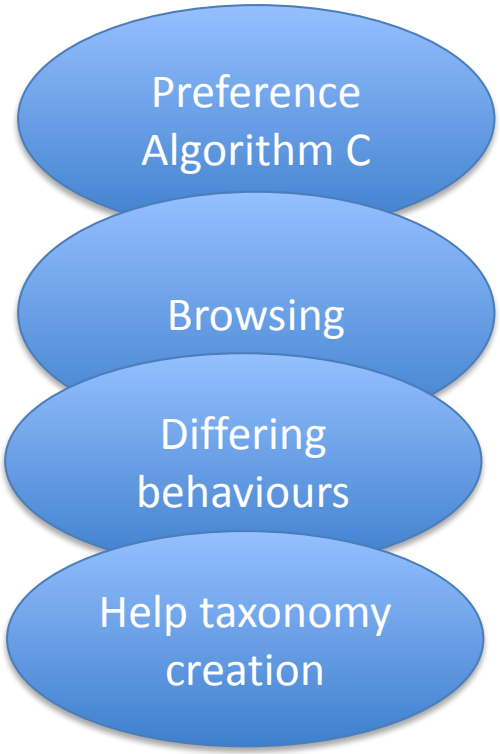
Co-occurring words
Algorithm A – Unigram
Algorithm B – Bigram
Algorithm C - Discriminatory



Interaction with touchscreens

RESULTS – QUESTION 3

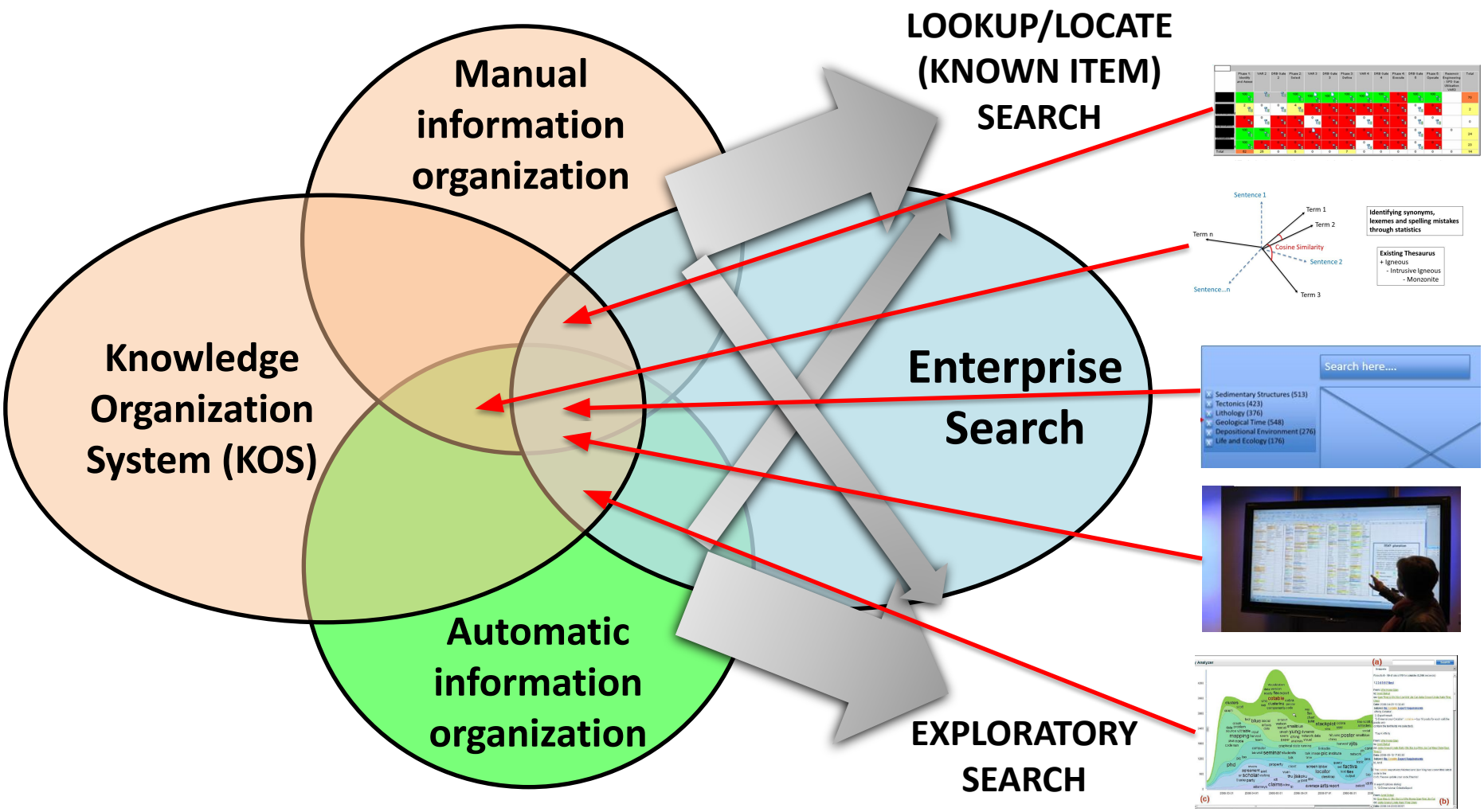
"Word associations highlighted new and unexpected terms… associated with the secondary keyword 'platform'. This surprising result led us to consider a new geological element which could impact our (exploration) opportunity" **[P32].**



"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]

"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 – ANIMATED TO EXPLAIN



CONCLUSION – BEST OF BOTH WORLDS

- Value in enterprises adopting *multi-methods* and *mixed methods* (with respect to manual and automated KO/KOS methods).
- Opportunities for enterprises to reconsider their strategies towards KO/KOS based on the theoretical model presented
- “Multi-lingual” corporate information professionals are more likely to facilitate innovations at the interfaces between disciplines

Thankyou for listening

- Email: p.h.cleverley@rgu.ac.uk
- Web: www.paulhcleverley.com