Innovations in Knowledge Organisation Conference – Chennai – 23 October 2017

www.ikoconference.org

Programme and Case Outlines

Conference Programme

9.30am **IKO 2017 opening keynote (plenary)**

Developing a Knowledge Organisation Community Response to the Post-Truth Information Age
David Clarke, FRSA

In 2016 the Oxford English Dictionary chose 'post-truth' as word of the year. Their decision was based on the proliferation of fake news stories and misinformation that accompanied both the US national election and the British EU referendum. On Earth Day in April 2017 thousands of people gathered in London for a March for Science, protesting the negative impact of post-truth culture and politics on science, research and education. How can the Knowledge Organization community engage with and respond to these issues? In this session, David Clarke will describe the problem space and comment on the issues from the perspective of knowledge organisation and information science. The session will conclude with 30 minutes for audience participation, including a discussion about how the KO community can get involved and contribute ideas for solutions.

10.55am Networking Break

11.15am IKO case studies - opening pitches for five case studies (plenary)

Maish Nichani (ISKO Singapore)
"Making Search Conversational: From Queries and Results Pages to Dynamic Conversations"

Patrick Lambe (ISKO Singapore)
"Developing Faceted Taxonomies from Knowledge Maps: A Case Study"

Rajesh Menon (Pingar)
"Proof of Concept Testing for Autoclassification: the Critical Step to Success"

Senthilvel Palanivelu, M.S. Muralidharan, S. Rajamani (SCOPE)
"Using Hybrid Machine+SME Methods to Develop a Faceted Taxonomy to Improve Search Effectiveness on a Large Corpus"

David Clarke (Synaptica)
"OASIS: Constructing knowledge bases around high resolution images using ontologies and Linked Data"

12.00 Lunch

1.00pm Case Study Café Round 1 - in-depth table discussions of 5 cases (35 minutes)

1.35pm Case Study Café Round 2 - in-depth table discussions of 5 cases (35 minutes)

2.10pm Case Study Café Round 3 - in-depth table discussions of 5 cases (35 minutes)

2.45pm Networking Break

3.05pm Table discussion followed by panel discussion - questions and issues for the panel

3.55pm IKO 2017 summary and closing address - Patrick Lambe, David Clarke, Maish Nichani

4.15pm Move to ICKOIM 2017 inaugural session, keynote and cultural programme

Maish Nichani (ISKO Singapore)

1. Conversational Search: From Queries and Results Pages to Dynamic Conversations

1. About the Case Organization
A government agency is responsible for the formulation and implementation of labour policies related to the workforce. It runs a contact centre which deals with many fact-based enquiries about the country's labour policies. Fact-based queries with predictable, precise answers are a strong candidate to automation via web interfaces, so that the contact centre staff can focus their time on more complex queries.

I will also show mini-demos of other organisations such as a Development Bank and StepTwo Designs, an intranet company based in Australia.

2. About the Challenge
The core challenge is how to deal with fact-based queries in a web interface using search.

We can observe two recent changes in web search behaviour:

- People are using long phrases when they ask questions (who won wimbledon last year)
- People are asking for specific fact-based answers (who won wimbledon last year; answer: andy murray)

Long phrases are good as they offer additional keywords and phrases that can be used to increase relevancy (e.g. what "last year" means can be inferred from the current date of the query; "won" and "Wimbledon" + date gives an easy answer). And being able to supply a specific answer is good as it eliminates the need for the user to inspect a list of search results.

However, people are also still using short fuzzy queries. Short fuzzy queries are harder to deal with in terms of predicting the specific intent and context of the user. Consider, for example, a fuzzy query such as: "swollen legs". A typical Google search results will list the top pages (ranked by other people's needs). While this might be good in most cases, it makes very hard if the person does not know what the available otions are, or if they represent an outlier, unusual case. Are swollen legs something to worry about? Can I treat them in general clinics? Can I buy a treatment off-the-shelf? Etc. While offering such options might be overwhelming for search on the public web, because people's contexts differ so widely, it may be useful in enterprise search. In enterprise search, if we have done our user research well, we have a better chance of accurately predicting the intent behind fuzzy queries simply by knowing more about the context and likely needs of the users.

We are currently testing a solution that combines the power of natural language search and conversational user interfaces to address this problem. This is the case I will share with you. We have completed parts of the solution but some other parts need to be done. The hope is that with this solution will people will self-serve and this will lower the demand on the contact centre.

3. What We Did
First, we powered the search bar to understand intent by training against a basket of natural

language utterances relevant to the intent. If the intent was fuzzy then we designed a dialog sequence with the user to ask them to clarify the available options, thus nudging them towards a richer query that could be understood and dealt with. At the end of the dialog (the fulfillment) we presented the user with super-relevant search results. The technologies used include natural language understanding, text anlaytics, machine learning, converational UI and a search stack.

4. Challenges and Lessons Learned

Trying to interpret language and intent is difficult. Mapping it to an ambiguous domain is even more challening. The lesson we learned is you can manage the complexity if the domain is narrow and the focus is on understanding a defined set of top (e.g. most common) user tasks (rather than the entire universe of possible tasks).

5. Impact and Benefits

We are currently testing the solution with the call centre reps. The feedback we are getting is that:

- It is easier to use
- It represents a faster way to get to the right answers
- It encourages them to explore (and discover) more as it is easy to ask follow-up questions

The expected business benefits will come from lower call volumes as more people will self-serve.

6. Next Steps

We need to do more work to integrate conversational UI with search. For example, how to transition from the search bar to a chat interface. From the organisation's perpective, they need to do their due diligence in investigating the new stack of technologies so that they can integrate it within their secure networks.

Patrick Lambe (ISKO Singapore)

2. Developing Faceted Taxonomies from Knowledge Maps

1. About the Case Organization
This was an international property development company.

2. About the Challenge
The company wanted to improve the way that its different teams collaborated and shared knowledge around major programmes and projects. Teams were used to working with shared folders with restricted access rights, and inconsistent or non-existent naming conventions for folders and files. Some had moved to SharePoint, but had simply transferred their shared folder structures to SharePoint document libraries. The lack of visibility into resources available in other departments meant that staff spent a lot of unnecessary time tracking down resources through colleagues they knew in other departments, or in reconstructing resources form scratch.

3. What You Did

- We conducted a knowledge audit for the client, building knowledge asset maps around the key activities of all the departments. This provided a "current state" set of descriptions of key knowledge and information assets across the whole company, associated with their key activities.
- We used a framework for describing knowledge assets that made it easy to decompose the knowledge maps into taxonomy facets describing activities, document types, project types, property types. The maps were created in an online system that made it easy for the departments to browse each others' maps and indicate which knowledge assets would be useful to have access to. We built the taxonomy around the knowledge assets that were identified for sharing, which (a) helped to focus the taxonomy on truly sharable assets and (b) did not have to describe the entire universe of informaton content.
- The innovative component of this project was the use of online knowledge maps to help focus the taxonomy on a shared knowledge base.

4. Challenges and Lessons Learned

- There was some resistance at the beginning to participating in the knowledge mapping workshops because of the attitude that each department was different, and had little potential to share. Once they were able to review their peers' knowledge maps, however, they realised that the potential for sharing was much greater than they had expected.
- Lack of initial widespread buy-in meant that participation was incomplete, and it took longer than expected to show the value from a shared taxonomy. We might have completed the project in a quicker period if we had targeted one division that saw the value of the project and used that as a demonstration project to show the value to other divisions.

5. Impact and Benefits

- The organisation now has a taxonomy that allows them to describe the information and knowledge resources they require to perform their work in a coordinated and

effective way across divisions.

- An unanticipated benefit from the knowledge mapping activity was the discovery that several divisions had a strong dependence on tacit knowledge embedded in the experience and historical knowledge of their staff. This meant that the taxonomy was able to describe areas of expertise for an expertise-finder system, and did not just describe information resources.
- The project was ultimately successful due to persistence and patience in the face of initial passive resistance from some parts of the organisation.

6. Next Steps

The organisation is now working on setting up a governance system to be able to maintain the taxonomy and network of experts in a sustainable fashion over time.

Rajesh Menon (Pingar)

3. Proof of Concept Testing for Autoclassification: the Critical Step to Success

About the Council of Europe
The Council of Europe (CoE) is a Strasbourg-based European organisation comprised of 47 member states. It was set up to promote democracy and protect human rights and the rule of law in Europe. Their content is primarily in English and French.

The CoE Challenge
CoE was looking for a document/content auto-categorization solution to improve the user experience finding documents amongst millions of files. Like any other organization, they find that manual tagging is time consuming and inconsistent and believed automating this could improve findability, but they needed to independently verify the accuracy of the automated tagging using some quantitative metrics.

What CoE and Pingar Did
In this Proof of Concept we tested how Pingar DiscoveryOne™ can handle auto-categorization. The CoE wanted a viable solution that can handle their existing set up, including the diversity in documents, departments and personnel as well as their existing vocabularies used as metadata.

We applied Pingar DiscoveryOne to a sample of CoE documents. DiscoveryOne used a feature we call KeyPhrases™ as well as five taxonomies co-adapted (within a restricted time period) by Pingar and CoE to automatically extract metadata from the documents.

The evaluation was conducted by collecting feedback from 13 CoE end-users in a controlled usability study, comparing with the automated tagging results. We will report on the evaluation metrics pre-agreed by Pingar and CoE and we will also cover insights from some other areas of interest.

Challenges and Lessons Learned
As we wanted to involve CoE in the whole process and take advantage of their domain expertise, we trained a team from the Information Life Cycle Division (the team who led the project from the CoE side) in using DiscoveryOne for adapting taxonomies for auto categorization. Pingar supported the CoE team along the way with answering questions and some further large-scale taxonomy adaptation. We also conducted the evaluation with CoE end users, receiving both quantitative and qualitative feedback.

The challenge with the PoC was to remain within the scope when both CoE and Pingar wanted to collect additional feedback on several aspects involved in content categorization. Both teams felt they could have benefited and achieved better results by allowing more time for several steps of the PoC (mostly adapting the taxonomies and setting up the evaluation environment).

On the other hand, we learnt a lot. Both teams need to see beyond the identified limitations of a time-limited pilot study and concentrate on the trends of the findings.

Impact and Benefits
Through this PoC, we were able to measure the preliminary success of DiscoveryOne on 3 predefined metrics (with the understanding that these can be improved further within a full project).

It also allowed CoE to understand the effort required for a full roll-out auto-categorization project as well as collecting end-user feedback on how appropriate their different vocabularies are for refining search.

What we did differently was to involve both the people who are the custodians of the vocabularies and the documents, but also the people who will be the end users of the auto-categorization project. This gave us a deep understanding of how the different types of metadata can benefit search users and the amount of effort required from both CoE and Pingar for a successful full project.

Next Steps
Pingar has conducted many PoCs, however due to the high level of engagement by the CoE team we collected a number of feature requests from them that we plan to introduce to DiscoveryOne.

As for running such pilot studies, we believe they are invaluable before scoping and successfully rolling out a full-on auto-categorization project. We will aim at reducing the limitations we identified during this PoC, as much as possible, mostly around time-demanding tasks by introducing more automation.

Senthilvel Palanivelu, M.S. Muralidharan, S. Rajamani (SCOPE)

4. Using Hybrid Machine+SME Methods to Develop a Faceted Taxonomy to Improve Search Effectiveness on a Large Corpus

1. About the Case Organization

ASTM International (ASTM), founded as the American Society for Testing and Materials, is a non-profit organization that develops and publishes approximately 12,000 technical standards, covering the procedures for testing and classification of materials. With over 30,000 members representing 135 countries, ASTM standards are used worldwide. Headquartered in West Conshohocken, USA, the organization serves as the administrator for the U.S. TAGs (United States Technical Advisory Group) and to numerous ISO/TCs (International Organization for Standardization/Technical Committee) and their subcommittees.

2. About the Challenge

What was the main objective, issue or problem we were using this approach to address?

- The search interface was linear with a flat list of topics, with users needing to drill down many results to select the most relevant documents for their research.
- Content spanning multiple subjects; required cross-domain search capabilities.

Prior to the case approach, how did the issue impact the business?

- Lack of structured navigation resulted in noisy search results. This negatively impacted the user experience and cascaded into poor sales conversion.

What size group/division was impacted by the case effort?

- ASTM provides standards, enterprise solutions, certification, proficiency testing, training and certifications for companies, standards bodies, government agencies, researchers, laboratories, and businesses ranging from Fortune 500 leaders to emerging startups. Industry verticals covered include metals, paints, plastics, textiles, petroleum, construction, energy, environment, consumer products, medical services, devices and electronics, advanced materials and emerging new industries — such as nanotechnology, additive manufacturing and industrial biotechnology.
- Some examples of users:
  - American Association of State Highway and Transportation Officials (AASHTO), is a Washington DC based standards setting body which publishes specifications, test protocols and guidelines used in highway design and construction throughout the United States.
  - Johnson & Johnson, headquartered in New Brunswick, N.J., a leading global manufacturer of sterile consumer products, pharmaceuticals and medical devices.
  - WhiteWater West Industries Ltd., headquartered Richmond, British Columbia, Canada - Designer and manufacturer of waterslides, multi-level water play structures.

3. What We Did

- Performed an extensive research on the corpus, user search logs, and other peer groups and concluded that a hierarchical and faceted taxonomy will improve the user experience.
- Deployed a team of Subject Matter Experts (SMEs) from the industry and the academia to develop taxonomies for 18 materials and 4 broad level facets or attributes of materials with a hierarchical structure – a faceted hierarchy.
  - For the 18 materials taxonomy, Scope suggested 4 cross-domain facets or filters – these were 1.Applications, 2.Properties & Measurements, 3.Process, and 4.Test Methods.
  - The keywords extracted from the documents are conceptually mapped to broader level concepts, viz., facets, which are the attributes of materials. For example, a material has a property that is subject to a process, which in turn is assessed through a test method used in the manufacture of products applied in specific industries. This is a sample semantic link between the materials taxonomy and facets.
  - As these facets are at a very broad level, further hierarchies are developed to drill down to a very specific attribute at a granular level. So, users can search or phrase queries on any facets of their interest and this will lead them to other related facets to explore further in their research process.

Technologies, Methods and Standards Used

- Scope has developed in-house process optimization tools for automatic clustering of concepts, machine –aided indexing and quality assurance.
- More relevant index terms were generated using Scope's proprietary automatic indexing tool, InDEXrTM, and further curated by SMEs. See details below.
  - NLP algorithms are used to extract noun phrases which could generally indicate concepts and these are used as candidate terms
  - Since these candidate terms may include all concepts discussed in a document, rules are framed to provide relevancy ranking based on the frequency and location heuristics.
  - Fuzzy logic rules are framed to facilitate some level of automatic clustering.
  - SMEs are used to validate and curate the automated output to remove irrelevant redundant concepts and include any key concepts missed through automation.
  - Supervised machine learning algorithms are used to improve the accuracy of automated output as feedback from SMEs are incorporated into the training sets.
- Scope has a resource base of certified taxonomists well versed in the NISO/ANSI standards for design, development and maintenance of taxonomies.
- Scope's technology team is well versed in W3C standards for machine readable formats for delivery of taxonomies/ thesauri in SKOS, RDF or OWL.

What was innovative about this effort?

- Scope first had a detailed discussion with the client on the specific pain points in their existing search interface – these are listed as below:
  - Users were being provided with a flat list of topics and there was no guided navigation or intuitive search, which resulted in users getting too many results for any topic search.

    - Content was spread across so many verticals and the content needed to be searched across verticals for concepts that could be common across several verticals.
- Scope did a complete corpus analysis and an analysis of the user search logs. The analysis included frequency analysis and conceptual labelling of keywords, clustering of closely related concepts, and analysis of how concepts were connected to each other through co-occurrence analysis. This helped to arrive at a theoretical structure of the knowledge model wherein concepts were grouped into very broad top level concepts, followed by development of hierarchies for each top level concept, and identification of the semantic relationships across these concepts.
- The entire process passed through a three-layer quality validation process with in-house SMEs doing the creation of the knowledge model, under the guidance of senior knowledge modelling and certified taxonomy experts. External experts from the academic stream and the industry for each domain did the final value addition.
- The ASTM team was initially apprehensive of how Scope's solution could be effective in enhancing the user experience, and improving the precision in navigation and targeted filtering to discover the most relevant results. To this end, Scope created a prototype demo interface to explain how the entire search experience could be enriched with Scope's proposed faceted taxonomy approach, which provided multiple entry points to the users, through taxonomy, facets and keywords. Scope demonstrated through the prototype, how users can drill down to the topics of their interest enabling contextual search, with minimal effort.

4. Challenges and Lessons Learned

Hurdles and barriers and how we covercame them

- As the complexity of the domains varied widely, services of external experts from academia and industry for some of the very specialized domains such as forensic science were employed
- The deployment of technology at the initial stages was slow and not very effective, due to the complexity of content and domains like forensic science.

Lessons Learned

- For knowledge intensive large-scale projects, it cannot be pure technological or SME solutions. We need to employ a judicious mix of techno-human solutions, enabling the technology to learn from the feedback provided by the SMEs and creating rules for expediting the automation processes without compromising on the high quality levels required.

Advice to another organization attempting a similar project

- Clients need to have a clear vision of their requirements, whether it is improving their existing products or creation of new products. They need to evaluate the suitability of off-the-shelf solutions where the degree of freedom to customize is minimal. Identifying a strategic partner who could convince and engage with the key sponsor in the organization will help in developing long-term relationships for innovative solutions.

5. Impact and Benefits

Business Benefits

- Minimized information retrieval time as there was a significant enhancement to the existing linear search platform by providing multiple navigation and search options to the users for knowledge discovery

- Compared to off-the-shelf vocabularies, Scope helped to create highly customized and content-specific taxonomies for 18 materials and 4 facets for ASTM within 10 months
- Created value-added features such as clustering documents by themes and interoperability by tagging keywords with internationally-accepted product codes
- Helped to extend the same knowledge model in multiple languages through translation of taxonomies and facets

Before approaching Scope, the client had explored many options such as off-the-shelf taxonomies and text mining tools that use advanced NLP algorithms. However, as these tools used general linguistic algorithms, it could not address the specific complexities of the various subject domains. Given that the verticals handled by the client are varied and each has unique characteristics and complexities, the client felt the need to develop a customized knowledge model for these verticals and then leverage these knowledge models for indexing documents.

Scope's detailed analysis of content and its ability to resource the specialist domain experts in each of the subject verticals for developing the knowledge model helped to provide high level of quality as required by a standards organization to satisfy their users.

The client now uses Scope's knowledge model for indexing third party content. Currently, Scope is working with ASTM to convert the taxonomy into a thesaurus.

Main reasons for success

- Scope used a clearly structured approach starting with corpus and user logs analysis, research into existing subject vocabularies and study of search interfaces in peer groups. This provided a solid foundation for the conceptual framework of the project.
- In this consultative engagement, Scope leveraged its strong network of SMEs, both in-house and from the academia and industry, to validate the algorithm based indexing and clustering. This approach of 'SME-assisted Automation' approach proved very effective.

6. Next Steps

Based on the experience from this project, Scope has conceptualized a consultative approach which include requirements analysis, proactive solutions, proof of concept demonstrations and agile project management using a team of SMEs and state of the art technology solutions.

Scope's smart content services can now provide integrated end-to-end offerings for knowledge modeling, maintenance, indexing and annotation. Furthermore, it can offer high-end ontology services for any domain.

Our client has extended their smart content framework in a phased manner to enrich the existing taxonomy as a thesaurus and moving towards a more robust ontological framework in the near future.

David Clarke (Synaptica)

5. OASIS: Constructing knowledge bases around high resolution images using ontologies and Linked Data

1. About the Case Organization
Synaptica produces software solutions for: building and managing taxonomies and crosswalks; designing and deploying knowledge organization systems; indexing and enriching content; and optimizing search, navigation and discovery.

2. About the Challenge
Visual images provide a valuable complement to textual information, but a vast amount of the information inside photographs, paintings, diagrams and drawings can be seen but not searched - it has been inaccessible to traditional query methods.

Many business applications could benefit from the ability to search inside images including: medical and scientific imagery, reconnaissance and intelligence, engineering and design, forensics and security, education and cultural heritage.

3. What We Did
Synaptica built a software system called OASIS that allows points and regions inside images to be highlighted and annotated. These visual features are then tagged using taxonomies and Knowledge Organization Systems. The software makes visual content searchable with pin point accuracy. It also promotes knowledge discovery as the application dynamically identifies features and related concepts as the user freely pans and zooms around an image.

The key technologies behind the solution are Linked Data and RDF graph databases. These allow users to connect to vast amounts of high-quality structured information in the Linked Open Data cloud, including authoritative Knowledge Organization Systems and ontologies. The extensive use of ISO and W3C standards and specifications ensures data portability and systems interoperability.

The fusion of several core technologies (high definition imagery, Linked Data, Knowledge Organisation Systems and semantic annotation) represents an innovative solution that opens up new opportunities for enriching visual content.

4. Challenges and Lessons Learned
Working with external data sources from the Linked Open Data cloud presents a number of challenges: (i) external data can be accessed by live queries to remote third-party servers, but these remote systems may not be able to provide adequate uptime availability or responsiveness; (ii) copies of external data can be ingested into local systems, but some datasets, such as DBPedia, may be too large to be accommodated on the available data storage; (iii) while graph databases out-perform relational databases at pattern-based queries, relational databases typically out-perform graph-databases at indexed or full text searches.

Synaptica responded to these challenges by building a flexible system that can simultaneously query data from any number of internal or external data stores. Low

volatility data of a manageable size can be ingested while high-volatility or very large datasets can be accessed on remote servers in real-time.

5. Impact and Benefits

The result of the effort is a robust and scalable general purpose toolset that can be used to build taxonomies, access external Linked Data, and annotate image content. By leveraging Linked Open Data, much of which is available without license fees, the time and cost to deploy knowledge organization systems can be greatly reduced.