INNOVATIONS
IN KNOWLEDGE
ORGANISATION

CASE STUDY CAFE – 15 JULY 2019 – PROGRAMME

Department of Library and Information Science at City, University of London

The IKO Case Study Café is a collaboration between the IKO Conference, an event of ISKO Singapore, and ISKO UK. It is an interactive event where you will have the chance to engage in in-depth discussions with a selection of case presenters on knowledge organisation case studies of your choice.

You will hear a series of pitches for 8 case studies in the main lecture theatre. The case presenters will then host a series of 3 x 30 minute table discussions in the foyer outside, and the adjacent rooms.

Please use the Case Study Navigator (overleaf) to locate your case studies of choice. This programme also contains the detailed case outlines to help you make your decisions.

14.10-15.10 – Case Study Pitches (Main Auditorium)

1. Using Knowledge Graphs to Model Standards and Business Processes in the Testing, Inspection and Certification Industry – Ian Davis, SGS
2. How Not to Implement Taxonomy and Search in O365: an Almost Disaster Story – Agnes Molnar, Search Explained
3. Danish National Police: Improving Search and Findability through Information Architecture, Governance and Taxonomy – Cecilie Rask, Danish National Police
4. Using Distributed Ledgers (AKA Blockchain) to Enable Trusted Exchange of Commercially Valuable Information across a Defence Consortium – Marcus Ralphs, Byzgen Limited
5. Cochrane: Use of Linked Data Descriptors of Primary Clinical Evidence to Support Meta-analysis and Facilitate Evidence Based Decisions on Healthcare Interventions – Julian Everett, Data Language
6. John Wiley: Developing a Molecular Biology and Biochemistry Taxonomy as a Non-SME – Niké Brown, John Wiley and Sons
7. Beyond Posting Counts: Giving Taxonomists a 360 Degree View of How Concepts are Being Applied to Content –Dave Clarke, Synaptica
8. Centre for Agriculture and Bioscience International: How a Single Knowledge Organisation Framework can Integrate Different Platforms and Activities –Anton Doroszenko, CABI

15.10-15.40 – Case Study Café Discussions Round 1 (Breakout Area)

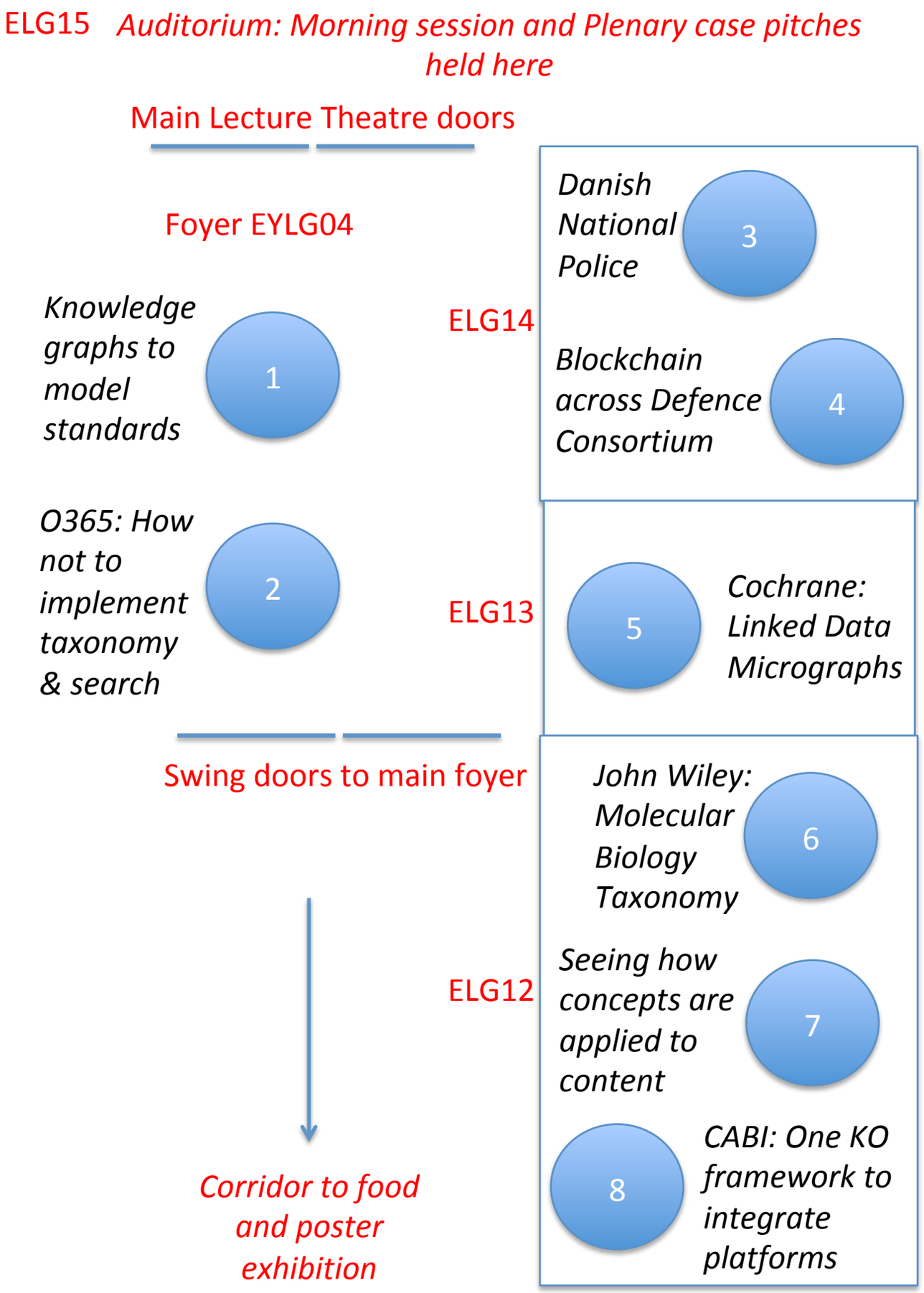
15.40-16.20 – Coffee Break

16.20-16.50 – Case Study Café Discussions Round 2 (Breakout Area)
16.50-17.20 – Case Study Café Discussions Round 3 (Breakout Area)
17.20-17.50 – Plenary Closing Discussions on Case Study Café (Main Auditorium)

Followed by Conference Reception

CASE STUDY NAVIGATOR

Use this navigator to locate the case studies of your choice



CASE 1: Using Knowledge Graphs to Model Standards and Business Processes in the Testing, Inspection and Certification Industry

Ian Davis, SGS
ian.davis@sgs.com

About the Case Organisation

SGS is the world's leading inspection, verification, testing and certification company. With more than 97,000 employees, we operate a network of over 2,600 offices and laboratories around the world. In the digital space, we market and sell thousands of services in 22 languages across numerous industries. Supporting our worldwide business operations are over 57 websites, a new ecommerce platform, a number of chatbots, and thousands of web and mobile apps.

About the Challenge

We aim to improve the way our customers and staff find and use information and knowledge across our digital estate. This involves implementing an SGS knowledge organisation system or 'KOS' incorporating aspects of a domain ontology with rich semantic structures and ultimately an enterprise-wide search solution supporting searching across all SGS digital content.

Starting with a one year proof of concept, we plan to build an SGS KOS to capture and connect varied information and data elements and to explicitly describe the relationships between them. We aim to model the underlying semantic structures of our domain. Our KOS ontology will support a data driven customer centric approach built on information exchange and knowledge development.

At a high level, we intend to:

- Lessen risk and compliance issues through tighter controls on digitally stored information
- Reduce duplication of information
- Strengthen and expand data portability and increase interoperability between systems
- Make explicit connections that enable customers to see related relevant information and offer the information they need when they need it
- Support content delivery in search, browse, personalisation and recommendations
- Empower AI using the logic and semantics in the knowledge model to support machine reasoning and improve search and categorisation
- Build smarter search and discovery applications by using logical dependencies
- Reduce costs and speed-up project deliverables by reusing a vast and rapidly growing library of public domain ontologies and taxonomies
- Simplify system integration by adopting open industry standard data models and portable data interchange formats

What We Have Done

We are at an early stage in our journey and are only a few months into a one year initial project phase.

To date we have:

- Audited what we have: websites, apps, supporting systems, vocabularies etc.
- Created a plan and a guidelines document to support the initiative
- Decided to:
    - Build to comply as fully as possible with ISO 25964 1 and 2
    - Use SKOS, Dublin Core and perhaps OWL

    - Add SGS relationships as needed between concepts
    - Use language that is user-friendly and concise
    - Develop a multi-lingual KOS with independent synonyms
- Identified stakeholders and audiences
- Agreed a communications plan
- Selected and licensed the Graphite system from Synaptica and begun work in the tool
- Created initial schemes; built a number of SGS specific associative relationships to link concepts; begun to add concepts to schemes and create relationships between them
- Sketched out integration targets and begun to plan integration

Challenges and Lessons Learned

Do not underestimate the time it takes to plan before acting. Create a broad plan describing:

- Objectives
- Existing landscape
- Construction principles and guidelines
- KOS scope and structure.

Do not rush into adding schemes, concepts and relationships without taking the time to think through the considerations and keeping track of what is being built. Expect legal and financial paperwork to take twice as long as you initially think and factor that in.

Think about the sources of truth before you build. Consider for each type of information where the source of truth will be and if it will be in your KOS or if the KOS will contain only a point of reference to the source of truth held elsewhere.

Consider exactly what concepts and types of information need to be modelled. Resist the temptation to model all aspects of a domain in detail unless they are needed. Ask why information needs to be in the KOS and what will it be used for.

Impact and Benefits

It is too early to see the impact and benefits of this work. We hope benefits will grow as the year progresses. However, we believe extensive business benefits will be seen after the first year of this initiative.

Next Steps

Over the next year we will complete our proof of concept. At the end of that period we will assess our progress and consider our next steps. This may involve closing down an unsuccessful project, developing the initiative at a relatively cautious organic rate or accelerating the project to realise our ambitions – in particular our challenging search aims.

CASE 2: How Not to Implement Taxonomy and Search in O365: an Almost Disaster Story

Agnes Molnar, Search Explained
agnes.molnar@searchexplained.com

About the Case Organisation

The organisation is a global manufacturing group headquartered in Western Europe, with over 30,000 employees. They decided to move their intranet portal to Office 365.

About the Challenge

The organization realised that they needed external help. They hired a couple of external consultants, including Search Explained, to help with their intranet and search strategy and implementation. However, they had been facing a few challenges:

- There were only a few months left until the global go-live.
- Up to that point, the business had not realised the full complexity of the migration and implementation issues.
- IT had the resources needed for the migration and roll out, however the team was not coordinated enough, neither did they have an overall migration strategy. As a result, they started to create content silos in Office 365, and their content started to become "mess in the cloud".
- Without a proper content and migration strategy in place, the organisation had no idea where they were with the actual migration process and what was still left to do.
- Due to the lack of proper training, the team expected "modern" and "smart" search to work much "smarter" by itself, and also to help with auto-tagging the content, and auto-identifying terms in the taxonomy.

With these pre-conditions the project was likely to be a disaster from the set-up, and we assessed that without a firm intervention the probability of success was close to zero.

What We Have Done

We helped the organisation to do a content inventory as well as a strategy with both short-term and long-term goals. We identified the must-have steps before the go-live, the steps to have a taxonomy professional in-place, and also created an action plan to follow.

To identify the priorities and governance needs, we used a content pyramid, with five layers of content – each needed layer would require different governance levels, content management and curation processes, as well as search configuration.

We also applied the "search-as-a-product" approach – we identified each phase with a version number and assigned goals and tasks to each. By using major and minor versions of the search rollout, we also created a draft communication plan to help the user adoption.

Challenges and Lessons Learned

The primary challenge was the tight schedule and the lack of team coordination. To overcome these challenges, the whole team had to be committed and success-oriented. Coordination was key.

My primary recommendation would be definitely not to wait until it's almost too late, as in this case. Implementing taxonomy and good search cannot happen overnight. Machine Learning, Artificial Intelligence, and "modern search" are all good and can help us, but only if the Information Architecture

is in place, the taxonomy is clean, and the governance, content management and curation processes are layered, aligned, and developed into a systematic migration plan.

Impact and Benefits

The primary benefit was definitely to meet the deadline and go live at the intended date. The team was successful, the stakeholders were satisfied. The "search-as-a-product" communication plan helped everyone to understand why search is not a one-time project, and why the organisation must keep investing in search enhancements.

Also, they identified a few inconsistencies in the taxonomy after the global roll-out, and with this approach they were able to correct them relatively easily.

Collecting feedback and having a solid two-way communication with the users also became part of their everyday lives.

Next Steps

The organisation has already stabilised the team and communications plan, as well as their action plan for the next one year and beyond. What they have to do is keep up the progress, and keep their team as coordinated as it is today – or even more.

During the next two years, more content will be migrated to Office 365 – we have to make sure that their current Information Architecture in Office 365 does not get siloed and disconnected again.

Regarding their search, the primary challenge for the next year will be to incorporate the new "modern search" as much as possible, and to identify where personalised results provide better results than the ones using classic ranking.

The company also decided to invest more into the taxonomy, in order to enhance both search and content navigation.

CASE 3: Danish National Police: Improving Search and Findability through Information Architecture, Governance and Taxonomy

Cecilie Rask, Danish National Police
cra018@politi.dk

About the Case Organisation

The Danish police is a national authority, whose purpose is to ensure the maintenance of public order and safety though prevention, investigation and prosecution of crime, and by dealing with conflicts. We have 15,000 employees and operate 24x7. In May 2017 the Danish Police launched a new intranet based on SharePoint 2013 with special focus on search functionality.

About the Challenge

In a police organisation, finding the exact right document and finding it fast is a must. We must ensure that officers on the street, investigators behind a desk and all other employees always have the right information to act on, across police districts and across subject areas.

We have 7 national centres as well as 14 districts including the Faroe Islands and Greenland. Some information applies nationally - this means it is relevant across all districts and to all employees - while other information is local and only applicable in specific districts. This presents rather complex requirements for targeting information on a need-to-have and good-to-have basis.

Findability was a real business problem on the old intranet. The search engine was outdated and there was next to no information architecture nationally or locally.

Often a national document was copied locally to ease the local findability. From a search as well as governance perspective, this was of course a nightmare practice, because instead of just one document, we ended up with 15 copies of the same document.

In addition, most documents were attached to news announcements. As a result, you would often have to know the year and date for the announcement, in order to find a particular document.

Another serious business issue was the more than 50 different document types. There was no consistency in how document types were used.

As a result, a common practice was to print documents and keep local, physical archives of paper. This of course also meant that the users often used old versions of a particular document or even referred to information that was no longer valid.

Search therefore became the #1 focus for the new intranet.

What We Have Done

From the very beginning, we prioritised building a solid foundation for search. We wanted to provide a functional working tool with a focus on search, governance and information architecture rather than a social intranet. In fact, it was much like building a house. We knew that the time invested in making the groundwork solid, would pay back over time.

We quickly realised that improving search would require more than a quick indexing of our content and the out-of–the-box search setup. Instead, we took a holistic approach to search, and identified a number of focus areas to support the overall search experience:

- We wanted to make the search user specific (by districts) – and at the same time make it possible to easily navigate and see content from other districts.
- We planned a new, common subject information architecture
- We defined a strict governance and information review life cycle
- We investigated options to present the national and local content in a more intuitive way for users, so they would only see content and documents relevant for them on a page (while they were still able to search across districts).

To achieve these goals we invested in search both from a people and time perspective. We invested time in finding best practices, attended courses, made connections to Microsoft MVP Agnes Molnar early in the planning phase.

The first brick was laid when the 50+ document types were replaced by a simpler document model with only seven well defined document types.

We decided to tag all content (both pages and documents) and defined a common and intuitive taxonomy, which would also serve as search filters.

Addressing the users' need for simplicity and predictability we developed a content setup using audience to display only the relevant content for a particular user (national content, which applies to all, and content from their own district). We coded it as a system feature to ensure uniformity across all

topics and sites.

At this point the search result page also started to take shape. In the beginning, we thought search filters would do the magic for us, but we soon realised that we could do so much more using the options in SharePoint search.

To fully understand the users' needs, we did workplace observations (e.g. with the duty officers who man the central stations and are responsible for all personnel on the streets). Based on this a new search results page was defined. It aimed at a simplistic display, information at a glance and solid filtering.

To ensure content quality and relevance, we decided on a 365 days content life cycle. We made an automatic, up-to-date overview of all content that needs to be renewed and we made it easy for editors to update/delete the outdated content directly from the overview page.

Challenges and Lessons Learned

Search is key and a little knowledge about search can take you a long way.

The consistent focus on search and findability has proven worthwhile, but it was only when we started to engage and develop the search in cooperation with users, that we really succeeded. Simply put, user engagement gives long-term value and enables better decisions.

It sounds trivial, but we had to educate users to trust search again. With the poor performance of the old search engine, many users had stopped searching altogether. With a continuous push throughout 2017, knowledge on how to search was channeled to users through local editors, simple written guides and videos.

We've centralised many functions and maintain a high degree of control. E.g. editors can't create sub sites and all pages are built the same way across subjects. While this can be perceived as very restrictive approach, we have found that it works to the benefit of the users in terms of a better overview across pages.

Impact and Benefits

Using the standard SharePoint 2013 standard functionality, we have transformed the out-of-box search to a business critical tool. Although we were not exactly considered heroes at first (it is not popular to change all of the information architecture, the search, and the document model all at once), the overall search experience has improved radically, and our users are generally positive.

We've created a sustainable content management process, simplified the document model and radically changed the search result page to improve users' general overview. All our content is tagged, which creates a good foundation for future search developments.

On the search results page sites and documents are separated:

- Sites are returned if the search term is either in the title or the front page is tagged with the term. This way we are able to ensure that if a user e.g. searches for 'fingerprint' he/she will see subject sites related to fingerprints e.g. forensics, forms of crime or asylum seekers.
- Documents are returned according to the standard ranking model but we display key metadata next to each document, so users can easily see document origin (national/local) and the document type. We have also placed a link to the site on which the document is stored. This way the user can easily navigate to the site and see all related documents and information in context.

Search is personalised and users only see results relevant to them (national content as well as content from their own district), but they can easily see content from other districts as well. And with a single click they can filter on document types to e.g. see only forms or actioncards.

Next Steps

We all know there is no such thing as a perfect intranet and our solution certainly continues to develop. Currently we are getting ready to roll out a new and more modern page design across all subject pages. We engage with users and do workplace observations, to understand how we can continuously improve the intranet functions. All development is now driven by user and business needs.

CASE 4: Using Distributed Ledgers (AKA Blockchain) to Enable Trusted Exchange of Commercially Valuable Information across a Defence Consortium

Marcus Ralphs, Byzgen Limited
marcus.ralphs@byzgen.com

About the Case Organisation

ByzGen is tech start-up based in London. Our data exchange platform employs peer-to-peer security protocols and Distributed Ledgers to securely share business processes and commercially valuable information, with verifiable states, across multiple, disparate, data management systems and fragmented data networks.

About the Challenge

In today's business environment, collaboration is often crucial to success. Frictionless sharing of valuable and sensitive information across traditional business and organisational boundaries, including supply chains, alliances and consortia will unlock the full commercial potential of collaboration; streamlining processes and reducing costs.

The problem is that organisations simply don't trust their commercially valuable data/information being shared externally and current data-sharing workarounds are clunky, expensive and often not very secure. This challenge is prevalent across the defence, security and aerospace sector and in complex manufacturing.

In late 2018, early 2019, ByzGen delivered a proof of concept, geared around a live prototype solution build, for the largest defence company in the UK; operating as the prime in a wide consortium of other defence and engineering firms. The programme we used as a Use Case to prove the concept was called TEMPEST; developing future stealth fighter aircraft.

What We Have Done

We were contracted to deliver proof that our platform could achieve the following:

- Trusted data exchange across a complex disparate network of multiple organisations
- Tamper-proof, real-time audit history
- Decentralised controls & originator ownership
- Provable data state

In order to prove these four capabilities, over a three-month period, we configured a bespoke version of our platform, including a fully operative user interface. Using live data sets, the platform was used to transfer commercially valuable Model Based Engineering Data from one organisation to the other.

Using AWS to host our decentralised network. Our proprietary middleware ran on a combination of

Hyperledger and Omniledger; one to handle the primary data and the other to manage encryption keys, access control lists and the audit history.

Challenges and Lessons Learned

The challenges we faced and subsequent lessons during this project fall into four categories:

- Understanding the business problem. It is quite easy to assume that a business problem exists. Unless we take the time to really understand the problem though, through detailed discussions and workshops with practitioners, the assumption will remain just that, and will prove difficult to translate into a reality.
- Prove business value. To prove business value, we must first understand the metrics by which that value should be measured: whether process efficiency, through speed and cost-based reduction, or in improved security, through provable improvements against the risk of network penetrations. Baseline data for the 'as is' must be sourced.
- Be careful not to fall into the 'interesting tech, but' category. Employing novel technology or using innovative technology in a new way, can easily become just the next shiny thing. Remaining focused on the capability and business improvements, both in terms of language and evidence, is critical in ensuring the core technology becomes more interesting than just what it actually does.
- Partnerships. As a start-up, partnering with a well established system integrator is extremely important. Not only do they provide confidence for the end client, but they also provide in-depth integration and implementation experience.

Impact and Benefits

Being a proof of concept, long term impacts and benefits are hard to define and quantify. However, the project successfully achieved the data exchange across a decentralised network and included a live 'under the bonnet' view of the data as it was encrypted, controlled and accessed, enabling the audit data to be compared real-time with what was actually happening in the platform.

We assess that the key impact and benefits going forward in this specific use case are likely to be:

- Improved ability to enforce standards
- Reduce operating costs through speeding up data-dependant design and engineering processes
- Accelerate the transition to cloud Storage
- Achieve complete data provenance
- Enable efficient and accurate real-time auditing

Next Steps

Over the last 12 months, ByzGen have successfully delivered three high-profile, fee earning, prototype builds. As a direct result of these projects, including this case study, we now have a live, fully deployable Trusted Data Exchange Platform.

In partnership with one of the largest system integrators in Europe we have co-built an end-to-end data exchange solution, focusing specifically on valuable intellectual property. This will be deployed on at least three customer networks (across defence, manufacturing and healthcare/pharma verticals), before the end of 2019.

CASE 5: Cochrane: Use of Linked Data Descriptors of Primary Clinical Evidence to Support Meta-analysis and Facilitate Evidence Based Decisions on Healthcare Interventions

Julian Everett, Data Language
julian.everett@datalanguage.com

About the Case Organisation

Cochrane's mission is to promote evidence-informed health decision-making by producing high-quality, relevant, accessible systematic reviews and other synthesised research evidence. Their work is internationally recognised as the benchmark for high-quality information about the effectiveness of health care.

About the Challenge

Historically, the systematic review production process at Cochrane was based on primary evidence collated via manual literature searches, which fed into a relatively complex workflow the ultimate output of which was a PDF document. Manual literature searching had both serious operational scalability limitations and also offered no reliable way of generating timely notifications of updates to the relevant primary evidence.

This resulted in a critical constraint on the volume of reviews that could be produced, and also created an unpredictable, long lead time in the publication of new versions of reviews that incorporated newly published additional primary evidence.

At the same time, the PDF output format created fundamental limitations on the accessibility of the data on which the review is based: a systematic review is essentially a meta-analysis result set which is then augmented with editorial explanation and interpretation. Cochrane had identified a significant number of product opportunities that they could not pursue as the review data was locked inside PDFs.

What We Have Done

Cochrane has a formal framework for encoding clinical questions in terms of the:

- Population impacted
- Interventions applied
- Comparators/Control against which the Interventions are being tested, and
- Outcomes for which the results are being measured.

We created a lightweight ontology that described this conceptual model and then populated it from the best fitting public domain controlled vocabularies for each property in the model. This included branches of SNOMED-CT, RxNorm, MedDRA, WHO ATC and MeSH. The approach allowed information specialists to represent the knowledge encoded in each review, study report and forest plot.

The result was a knowledge graph for clinical evidence that could queried by other review authors and contributors, review consumers such as guideline developers, semantic search tools and external partner APIs.

The knowledge graph was populated by a combination of crowd/citizen scientist, machine and subject matter expert inputs, and curated by information specialists. To do this, they used a portable suite of web annotation tools (which could be dropped into existing web applications) and a quality assurance workbench, and the results were then made generally available within the organisation via a set of APIs and a clinical evidence search interface.

The same services feed into a target product pipeline based on the ResearchObjects model (http://www.researchobject.org/) that defers the packaging of the data and content until the delivery phase, allowing it to be tailored to the specific needs of the product and user contexts, be that PDF, JSON or anything else.

Challenges and Lessons Learned

- Vocabulary selection, management and curation has been, and continues to be, arguably the most important challenge, as it creates the information management foundation that enables everything else. Adding missing terms, reconciling synonyms, removing duplicate terms, deprecating terms and reviewing new updates of third party vocabularies has become a full time role for the information specialists who previously conducted manual literature searches. Several other organisations in the sector are now also attempting to tackle this challenge, including NHS Digital, and a community-led collaboration seems by far the most likely way to build a sustainably manageable resource for the benefit of all.

- The need for different workflows which address annotation of the archive of existing published content versus the creation of new content was also a very significant challenge, the latter having very substantial change management implications for core editorial business processes. This was addressed by tackling vertical slices of the archive first (with no impact to existing ways of working), demonstrating the solutions this enabled, generating stakeholder buy-in, and only then looking to integrate into existing workflows.

- The information management tools were designed as a widget library to be dropped into any web application. Unexpectedly this has turned out to be most useful to external partners who wanted to use the tools, whereas most internal consumers ended up developing bespoke UI components that completely suited the user experience of their host applications.

Impact and Benefits

In addition to the benefits to the content production process already described above, further unforeseen benefits were created via the ability to offer tools and services to external partners. Via the design of loosely coupled services and associated tools which discretely encapsulated each of its core business capabilities, we were able to offer not only the content it produced but also the tools used to create that content to its audience and partners. This is changing its role in the healthcare sector from being a content-only company to a combined content and services organisation.

Next Steps

The next steps for Cochrane are to complete the roll-out into the review production tools, to complete the evidence surveillance and notifications services, and to seek other partners with whom to develop a community-owned terminology curation capability.

CASE 6: John Wiley: Developing a Molecular Biology and Biochemistry Taxonomy as a Non-SME

Niké Brown, John Wiley and Sons
nbrown@wiley.com

About the Case Organisation

John Wiley and Sons is a global academic publishing company, publishing over 800 journals on behalf of approximately 600 scientific and specialist societies whose content, in some cases, goes as far back as the 19th century. Wiley's continuing objective is to deliver content to educators, learners and professionals in new and innovative ways.

About the Challenge

One of the most important features to add value to online content is taxonomic classification, which enables faceted search and discoverability within and across rich and deep content collections.

The Content Enrichment team is small and does not possess expert knowledge of all the subject domains Wiley publishes content across. This obviously hinders efforts to provide taxonomies for societies and their publications. Many societies are fully aware of taxonomies and their benefits, but few want to take on the work to create their own due to perceived time and resource limitations.

In some areas there are taxonomies which have been built by external providers, but they are often either not available for reuse or not suitable for the specific content published by Wiley.

Lack of taxonomies for content classification means that the added value features are not available to the society portals for their authors and readers. This poses a serious commercial problem when a society's journal publishing contract comes up for renewal, or when the company is courting a new society to join Wiley.

What We Have Done

Wiley had been approached by an external vendor, Scope, with the idea of building a taxonomy from scratch as a Pilot; taxonomy construction is one of the services they offer, and they have in-house SMEs covering a range of knowledge domains. Scope already work on content classification for the Wiley, so are well known to the company.

A pilot was set up to build a thesaurus covering the subject domain of molecular biochemistry, tailored around the content areas published by three specific societies. The content enrichment department would pay for the pilot and present it to the societies gratis. If the pilot and model developed during this project was successful, it could then be offered to more societies as a solution to the taxonomy creation problem (if one was not pre-existing, of course).

Scope used the following technology and standards to create the thesaurus, agreed between Scope and Wiley prior to the start of the pilot.

- Synaptica Knowledge Management System for corpus analysis.
- Industry standards ANSI/NISO Z39.19 and ISO 25964

The finished thesaurus would be delivered to Wiley in SKOS format for import into Wiley's thesaurus management software. A thesaurus was commissioned as we wanted related terms included as well as hierarchical taxonomies.

To help with validating that the work from Scope, an internal Wiley editorial team of three people (later down to two people) who are SMEs in molecular biochemistry (but not in thesaurus construction).

The approach taken was to divide the overall subject area into 11 domains, and then to build taxonomies for each one, ie taking a 'top-down' approach. Each domain was then submitted to the editorial team for validation.

When all domains were approved, Scope built up the related terms between the taxonomies (with some RTs within the taxonomies), thus creating a thesaurus.

This work was done in Excel spreadsheets. Scope selected terms based on their frequency in the corpus of documents supplied to them by Wiley, and then submitted them to Wiley for validation. This was an iterative process, refining the term selection until Wiley was happy with the results.

The document corpus supplied to Scope was filtered in order to remove bibliographies and other extraneous material, and to only include primary research articles. The approach to concept extraction taken by Scope was the following:

- Analysis of the input corpus content plus author keywords, industry standards and inputs from the subject expert inputs.
- Reference to trusted repositories such as iMedPub, MeSH and PLOS Thesaurus etc.
- The indexing platform used by Scope uses a blend of natural language processing (NLP), location heuristics and statistical algorithms for extracting keywords. This sort of platform automatically includes stop words.
- The resulting candidate terms were normalised to ensure correct forms for terms. They were then combined and de-duped before frequency of terms was assessed.
- A number of quality calculations were applied to the terms to achieve standards of accuracy, comprehensiveness, hierarchical relationship and taxonomy development quality.

This approach to thesaurus construction has never been done before by the company – if successful, it would provide a model for future thesaurus development. One of the main objectives is to create a subject-specific thesaurus which is owned and maintained by Wiley, and which can be reused as required. This pilot was initiated for three specific societies and their content, but two more societies have already shown interest in using it for their content classification.

Challenges and Lessons Learned

The main challenge was the understanding between the editors and Scope as to what terms would be appropriate for the thesaurus – to begin with, Scope's selection was textbook and did not cover more specific content as expected by the editors.

Another issue that arose was an exact understanding of what a thesaurus is and expectations around what was to be delivered – Scope were rightly adhering to the standards, but this proved restrictive until I clarified the approach to be taken, ie broad and shallow, not rigidly supplying levels if they did not naturally form. To overcome this, we instigated two regular calls a week to discuss these issues rather than relying on email exchanges. This proved highly effective in quickly resolving misunderstandings or confusion and the teams were able to move on more efficiently.

Next time, I would ensure that a more exact brief is supplied, plus a better outline to the project and how communications are handled. The highly collaborative approach we took worked really well but I also appreciate we were very fortunate to having a highly engaged editorial team involved in the project.

Also, not to be underestimated is how people think about a thesaurus, and their concerns about it not being done 'correctly' or being totally perfect. This is Phase 1 of the thesaurus and emphasising this, along with the reassurance that 'mistakes' can be corrected easily, is essential in helping those involved in relaxing when attempting to include or discard concepts. A pragmatic approach needs to be adapted – if the content does not exist to be classified in a particular subject area, (and the thesaurus is being designed for this purpose, of course), then don't bother creating terms for it!

Impact and Benefits

The business benefits from taking this approach are to overcome reluctance both internally and externally to having content classified – thesaurus (or taxonomy) construction is often viewed as 'too difficult' to do as well as time-consuming and resource-hungry. But academic societies, researchers and authors also want their articles to be visible and discoverable which requires the help of taxonomies. The aim here is to use the pilot as proof that taxonomies can be built relatively quickly, cheaply and can be reused and adapted for more societies by taking a subject domain approach as opposed to being built specifically for particular journals.

The success of this initiative is still to be measured, but it shows great promise and gives Wiley a well-defined content enrichment model to offer to both internal content teams and external societies. There will also be opportunities to offer the thesauri to interested external parties – many of the subject areas currently being considered do not have pre-existing thesauri created either by Wiley or anyone else.

Next Steps

- Further develop the thesaurus adding in synonyms and antonyms. The RTs will also need validating.
- Test the thesaurus by classifying content to check the terms – are they all used? Some but not others? Where are the gaps?
- Demo the thesaurus to the interested parties and get feedback from them. Initial viewings have already received very positive feedback.

CASE 7: Beyond Posting Counts: Giving Taxonomists a 360 Degree View of How Concepts are Being Applied to Content

Dave Clarke, Synaptica
dave.clarke@synaptica.com

About the Case Organisation

Synaptica provides enterprise taxonomy and ontology management software tools and professional services. Our mission is to help people organise, categorise, and discover the knowledge in their enterprise. We are located in Colorado with a distributed team of software developers and KO/KM consultants in the US, Europe and Asia. The case organisation is a multimedia corporation producing and distributing interactive titles to a global audience.

About the Challenge

To be effective with the long-term management of taxonomies, taxonomists need to know how terms are being applied to content. Simply put, a concept that has been used thousands of times is one that one should think carefully about modifying or may need splitting into more granular concepts. Conversely, a concept that has seldom been used may be one that needs to be withdrawn or merged with a more popular concept.

Traditional library science developed a methodology known as 'Posting Counts' in which a metric is displayed within a thesaurus or taxonomy system to quantify the number of documents or pages each

concept is indexed to.

In 2019 Synaptica worked with a client who needed to push the envelope further and provide detailed information about tagged content directly within the taxonomy management system.

The initial deployment of the solution will be used by a team of taxonomists and by thousands of end-users of an enterprise Wiki linked to the taxonomy system.

What We Have Done

The solution we will review in this case study used ontologies to create content description profiles and bi-directional APIs so that when a page of content is tagged with a concept it also creates a bibliographic record about that page within the taxonomy management system, as well as a link back to the concept.

The resulting solution provides users with an innovative ability to review qualitative as well as quantitative details about how any concept is being used. Links let taxonomists jump from concepts to content summaries and from there on to related concepts and related content without leaving the taxonomy management system.

Challenges and Lessons Learned

The use-case presents challenges of scale and change management. A taxonomy of a thousand concepts may easily be indexed to millions of pages, which then need to be made searchable and/or browsable within the taxonomy management system. What happens when content pages are deleted, renamed, or moved?

The ability to track and propagate updates from the content management system to the taxonomy management system is complex and subject to some inherent constraints of the content management software.

Impact and Benefits

The developed solution has been functionally tested and performance tested at high scale through an extensive proof-of-concept trial. The production deployment is scheduled for Q3 2019, so the full business impact has yet to be assessed.

Two major benefits are anticipated once the system has been scaled-up across enterprise taxonomies and content: (1) by providing taxonomists with qualitative as well as quantitative details about how concepts are being used, the taxonomy curation and governance process will be improved delivering greater process agility and semantic relevance; (2) by combining taxonomies (Knowledge Organization Systems) and content metadata within a single Linked Data triple store, a knowledge graph is formed that can support complex pattern queries and the ability to deliver actionable insights.

Next Steps

The client for whom this solution was developed will go into production deployment later this year.

The software solution has been designed as a set of end-user configurable profiles, plugins and APIs, thereby allowing it to be adapted for rapid deployment to meet the specific needs of other clients.

CASE 8: Centre for Agriculture and Bioscience International: How a Single Knowledge Organisation Framework can Integrate Different Platforms and Activities

Anton Doroszenko, CABI
a.doroszenko@cabi.org

About the Case Organisation

CABI is an international not-for-profit organisation working in over 40 countries to help solve problems in agriculture and the environment. We create and curate related knowledge resources and we work with others to apply that knowledge. CABI serves an extremely wide constituency, including academic researchers, librarians and farmers, covering a very broad range of topics. The organisation currently has over 500 staff in 12 countries.

About the Challenge

The challenge faced by CABI is firmly rooted in its long history in publishing and research. CABI began curating the scientific literature covering life sciences and related topics in 1910. By the 1960s there were a dozen bureaux and institutes in Europe specialising in different subject areas, such as entomology, crop and animal science, forestry, engineering, rural economics, parasitology, and mycology. We also had regional offices coordinating research projects on four continents.

The shift to electronic products, particularly since the 1990s, exacerbated previous diverse approaches to metadata management. Frankly, the CAB Thesaurus at the turn of the 21st Century did not have the breath of coverage required for the new electronic age. At that time terminologies were scattered all over the place, including in specialist databases created by our research institutes as well as in numerous validation files used for publication production processes. Additionally, particular projects on tight delivery schedules created their own taxonomies to satisfy the demands of project donors.

This fragmentation of taxonomies was having a detrimental impact on coordinating existing products and in planning new ones. A unified approach was key to changing this situation.

What We Have Done

In 1999, the last print edition, our thesaurus had only 63K terms in English, with translations into Spanish and Portuguese. Shortly after that we started using dedicated thesaurus management software. By 2008 it had increased, by purely manual processes, to 81K terms in English with no increase in the translations. But there remained enormous gaps in subject coverage, particularly organism names.

When I took over the management of the thesaurus in 2008 it was decided to mine existing information sources available within the CABI publishing ecosystem. We felt, quite reasonably, that these sources could be relied on for acceptable data quality. For example, we had synonym files used to add preferred terms to the indexing of our bibliographic databases. Additionally, we went to selected data sources outside CABI.

Mining of structured databases produced by CABI research staff also provided the opportunity to add ontological-style relationships to the thesaurus in addition to the traditional hierarchical structure. Currently, we have Crop Plant <-> Harvested Product, Disease Agent <-> Disease Name, Biocontrol Agent <-> Host, and Disease Vector <-> Vectored Agent.

Several of these mined data resources were simple text files, which were easy to handle. Other sources were specialist databases requiring restructuring and reformatting into import files. The thesaurus management software made it easy to import large amounts of structured data into the

thesaurus using simple text files without the need for IT support.

Therefore, the thesaurus has grown from approximately 200,000 terms in three languages in 2008 to 2.9 million terms in eleven languages today. Importantly, this growth has come without jeopardising data quality. It was never just about quantity of terms.

Challenges and Lessons Learned

- Long-established products create their own inertia. Some CABI projects that created their own taxonomies out of necessity many years ago are finding that it is now difficult to align these to CAB Thesaurus. It took some persuasion for project managers to change course because of the amount of work required. However, it is being done because of the advantages of using the same source of terminology across all products and projects. The lesson is not to 'do your own thing' in future.

- Repurposing data used by other CABI systems threw up a few inconsistencies. For example, some files used to enhance metadata in our bibliographic databases deliberately included orthographic variants in order to finally output cleaned data into products. But it wasn't easy to identify those spelling errors because the files were so large. These errors added to the thesaurus have taken time to correct.

- Despite the content gap initially being huge we didn't fall into the temptation of scraping the Internet. That risked gathering huge amounts of poor-quality data. To illustrate, manual checking of data taken from a web site ostensibly created by domain experts revealed that 30% of the organism names were incorrect. Half of these were spelling errors; the other half were due to use of out-of-date taxonomy. In any case, almost no external data source is in a format that can be easily utilised.

- We learned that manual checking is by far the best way to verify content. While the temptation is to use automated processes even specialist databases in the same domain frequently disagree. Disentangling these discrepancies required human intervention that artificial intelligence would struggle to resolve. At the moment it is just too complex for a machine to determine the correct resolution in most cases.

- Ontological-style relationships are proving their worth for many projects. The added value of these relationships is where further knowledge can be embedded. It adds to the contextualisation provided by the traditional thesaurus hierarchical structure.

- Using the correct tools enabled rapid growth of the thesaurus. Not having to rely on IT expertise meant that the thesaurus team had complete control of the thesaurus and were not delayed by the requirement for scheduling of IT work.

Impact and Benefits

The positive impact of standardising the terminology across all of CABI's publishing products has been gradual. It has taken several years. Nevertheless, persistence has paid off. By creating our own thesaurus, we are able to create products unique to CABI with several added-value features not available elsewhere. Sales growth seems to corroborate that assertion. Project management costs are also reduced by creating one taxonomy and making it available from one source maintained by one team.

Customers prefer having all the information they require in one place. For example, we can now link compendium datasheets with relevant bibliographic records, book chapters and full text content and be almost certain they are all concerned with the same topic. Creating a data repository for all of CABI's content allows customised, integrated products to be created on the fly. When pulling in content from outside CABI there is more confidence that this content is relevant as well.

A more comprehensive and steadily improving thesaurus has allowed CABI to fill gaps and correct errors in metadata. This included approximately 2.5 million bibliographic records published in the pre-electronic age when they were converted into electronic form. Annual sweeps of our entire content have kept the metadata current and have significantly enhanced relevance and recall in searches by our customers.

Next Steps

It is anticipated that CAB Thesaurus content will continue to grow, but at a slower pace over the next five years. However, the thesaurus as it now stands is robust enough to create new tools and services. Our particular focus will be on enhanced (under the hood) and guided (user assisted) search tools, services based on the thesaurus via APIs, fully exploiting open data using URIs, enhanced ontological relationships, and enhanced use of translated terms to increase product usage in non-English markets.