BREAKING DOWN KNOWLEDGE GRAPHS, TAXONOMIES, CLASSIFICATION AND ONTOLOGIES
VIDEO TRANSCRIPT

Hi everyone and welcome to the next webinar in our special webinar series on knowledge management - breaking down graphs, taxonomies, classification and ontologies. Just want to cover a few logistics before we get started. Everyone has been muted to eliminate any background noise and if you have any questions during our presentation, there is a Q&A box on your screen, so you feel free to ask them there and we will cover those questions at the end. And now I'd like to introduce today Speaker: Paul Nelson was an early pioneer in the field of text retrieval and has worked on search engines for over 30 years. He was the architect and inventor of retrieval Ware now owned by Microsoft Corp. Paul served as the chief architect at Search Technologies until their acquisition by Accenture 2017 and he is now a master technical architect with Accenture Applied Intelligence where he continues to provide architectural oversight design technology research and training. And I'll pass it over to you Paul.

Thanks Susanne. We have a packed number of slides so will get right into it. We are part of Accenture and Accenture is one of the largest knowledge repositories in the whole world. Because we have just so many people and we do so many different projects and you know it's really across every possible technology in across all industries. Our group is inside of Accenture and is called the Search and Content Analytics group. We're focused pretty much on search engines, unstructured content, natural language processing, and as our subject is today, knowledge management. So you know for these series of webinars that we've been doing this fall, we're really trying to focus on the top 5 challenges. The first is the volume of knowledge. Knowledge is everywhere. How do you acquire that knowledge from all over your organization across many silos and make that available you know a consistent way to all your employees. The second is when you need to have curated knowledge versus organic knowledge that simply arises from the way that you do business. The third, which is the subject of today, is graphs, taxonomies, classification ontologies. And by graphs, we mean knowledge graphs - how do we sort of navigate what these things are and when do you need them and when should you use them. Context and linkage the unsung heroes of KM that will be our next webinar, and then finally how do we justify measure and prioritize knowledge management to ensure that it's achieving the business value.

And of course today will be talking about graphs technology, taxonomy and classifications on top and all of that. So the reason why I wanted to give this particular webinar an hour is because the terminology is kind of dense. What is a taxonomy really? What is an ontology really? And you know I go into so many meetings where these terms get tossed about and a lot of folks say, “We definitely need a taxonomy.” And I'm like “Well what for?” and so I just thought that this webinar might be helpful just introduce the terminology and describe exactly what it is from our perspective. And then describe you know, what does this mean for an actual knowledge management system - when would you use these things how do they help? You know what scenarios do they make sense and one of the things. I think that's kind of a theme is a lot of terminologies come from the world of academic academia, and you know the world of academia may not be that great for the world of business. When you are doing it, taxonomy for example, for a library and you know those books, they last for ever.

And you do it perhaps for an archive like the National Archives where you know that archival material that last forever these are long term institutions with very stable you know understandings but when you translate that into the business world where every month you know last month's presentation is already out of date you know maybe a lot of these things don't apply and that's why I just want to be super clear about what all of these items are and you know when do they apply for real knowledge management system you know for real business.

Ok so let's start with taxonomies you know a taxonomy is really just a classification hierarchy so if course I say that it's really just categories. You know taxonomy classification hierarchy taxonomy is are basically all the same so somebody thought you know I have this massive collection of documents wouldn't it make sense if I just like categorize them somehow and you know. That this would just organize that big pot of documents in a way that would you know be helpful and then maybe in the subcategories as well so I have categories and I have some categories and now we start having this sort of category hierarchy or classification hierarchy and of course you could also have crossed categories you know multiple different ways different dimensions that you could sort of organize your big collection of documents into these subsets. So you know what is a taxonomy exactly you know it's also known as just categories or classification hierarchies you know like in a library the implications are that the document appears in one place now of course that's not always the case there's lots of cases where you would have documents with you know multiple points you know in your taxonomy but the traditional sense you know like an animal like a dog or a platypus only occurs one place in that taxonomic classification of animals or a book write a book only it is on one shelf in the library because you only have one copy of the book you have to put it someplace where people can find it so that that's the implication but in practice with electronic documents they can appear you know many places in the hierarchy typically they're multi-level typically information special specialists who maintain it and it's typically a classification of the whole document you know not just a reference within the document.

And you know the advantages you can search over a smaller set of documents which is inherently more accurate right searching over you know a 1000 documents is way more accurate than searching over a 1000000 documents and people can subscribe to categories so that's really nice that if I have a subject category called block chain and I want to subscribe to that to start seeing new things that come in about block chain. It adds some organization and you know we can do some machine learning to automatically assign things to categories you need a bunch of examples but once you have those examples the machine writing is so much research around that it can do it pretty accurately. So I think the problem is that category taxonomy doesn't scale for most businesses because the world is complicated you ultimately know you get an explosion of tags there's a lot of ambiguity. You get these documents that fall into that like these gray areas and corporate change you know corporations get restructured you know our understanding of the world gets restructured to things just change too quickly. So generally, I only like to use taxonomies when they are applied to an actual business function like a business group or corporate function or manufacturing or when they're organized in a well defined an industry standard process like drug discovery or construction projects or legal cases or when someone else is maintaining the taxonomy then I think it makes a lot more sense, so that you personally don't have to invest to have a bunch of data scientists to maintain your taxonomy. So if you really want to have one you know generally the smaller the better in focus groups like if you assign a tag to a document how people know that they have to place that tag on that document and when they go to search and use that tag will they really know what that tag means so you know to test it and post focus groups create your taxonomy, ask 10 people to categorize 10 documents as 10 people you know what do these mean.

But you know they're only as good as the quality and you know what I've seen is that a lot of people do tags, and then 3 years later nobody ever uses them because they're just not applied accurately. People don't know that they exist; when people know that they exist, they don't know what they mean. You know it's just the sort of thing that a lot of people add tags, a lot of people add taxonomies, and they just don't get used that much. So if you want to use them, consider using publicly maintained definition like a Wikipedia. So if I tag something with a java then I know exactly what it means. Is that the coffee? Is at the language? You can have a Wikipedia topic to actually define what the tag is and then identify the scope of the tag. Does it represent the whole document or is it just a mention within the document. Consider creating a primary and secondary tag.

Let's move on to ontologies this is you know such a deep subject and when you lock look at the world of like what is academic ontology discussions of these are just such. I mean it is truly an existential question like we say all that's of an existence a question we just mean that it's a hard question. What is the existence of a property, what is existence, what does it mean to be, how or what is qualitative, what is quality, what is a physical object? Are entities objects? What are the states various modes of being? I just love these questions and you know I just kind of want to spend my whole life just to answer these questions like I imagine a lot of philosophers do. But what is really an ontology like for knowledge management? And I think when people talk about ontology as for knowledge management, they're mostly talking about domain ontologies. Let’s define the world we need to talk about. What are the entities? What are the properties? What are the relationships of the entities organized in a Taxonomy?

I mean what are the properties of the entity, what are the relationships between things, how are things are related, how are the relationships identified, what do you care about that in your world and what's not in your world. And then how are these things described so that you can correlate sort of narrowed down that when you talk about.

It means one thing rather than like 3 possible different things. OK, so that helps a little but really practically speaking, for knowledge management getting right down to the nitty gritty. Really ontologies are about objects fields values and foreign keys. So what are the business objects you care about? OK, customers, employees, business groups, documents, presentations. What are those objects? How are they identified? Are they in the categories by subject, by geography, by product line? So all that together gives us the ontology for your domain, ontology for your world. And then when we talk about properties, those objects, do they have fields? What are the values in those fields? What are your date formats? What are your file formats? Are you using types? Are you using email addresses? How are things and people identified? So all that as part of an ontology as well as the foreign keys and how things are related to each other. Just a couple examples. The 1st examples from Dublin Core. And so a Dublin Core is an ontology. It's fairly lightweight. It's the domain is web pages which are documents and so these are the sorts of things like publisher, contributor, creator, dates coverage, relationships, sources, and subjects. And you can see where in Dublin Core they've identified recommended values like the language code, the date structure, ISBN identifiers, and all sorts of things so that’s Dublin Core. That's an example of an ontology right there. Another example this is kind of more of a business example. These would be the entities in your business: employees, organizational groups, cost centers, office locations, Internet destinations. Then how they're related, how they are grouped, and you can see some of these have higher keys in them like parent child subgroups, like supervisor and organizational chart hierarchies.

And so how are these items related and the metadata fields for each of these items, and what sorts of values you can have, and do you have essentially an enumerated type of different types of groups or do you have an enumerated type of you know different levels? Maybe you have 8 levels and so then you know the level field will be a number 1 to 8. And so that together gives you the ontology for the domain, ontology for your business world. So just some recommendations ontologies. 1st, don't overdo it but it's the same time as don't under do it. A little bit of planning, trying to avoid collisions between entities, sticking to some basic standards like date formats, number formats, documented like write it down, and even if it's just fairly a lightweight like a 5-page document which describes your ontology, that will just help. As you build out your knowledge management system, make sure that all the different parts of your knowledge management system across your organization play together well. Your structures extension of all this is where JSON is very helpful and then plan for relevancy. If you're going to be searching over all the items in your in your world, then you just want to plan a little bit ahead so that you can know how it fits with the search engine for relevancy scoring and especially when you need like external data like sales volume, organizational level, seniority, popularity. Those sorts of things really help with relevancy for a search engine, so make sure you have fields and those sort of data stores organized a little bit before you begin.

OK, so there're 2 more sort of subjects. One is controlled vocabulary areas and the other is the final one is knowledge graphs. And we'll try and finish those up in the next say 10 minutes or so. So what is a control vocabulary? You know it's useful to have a common way to say things because there are many different meanings in different domains and having a controlled vocabulary helps. When I say this thing like the word “spleen” I know exactly what I'm talking about. And it's just very important for some domains like medical, nuclear power, things that are like critical domains where you really want to know that when you say something, you mean something.

And so an example of all controlled vocabulary is the medical subject headings from the US National Library of Medicine, which is all of medicine used by everybody. So this is a little more flexible it's words and synonyms sets essentially a controlled vocabulary for English with descriptions. But we do have these sort of less controlled vocabularies which I would recommend like the source just an ordinary thesaurus like you would have to identify a cash and liquid assets and legal tender. Maybe those things are very similar or company or conference call and zoom, teams meeting, you know those sorts of things just help bridge that gap between what a user types and what the doctrine is. However, as it turns out, your own networks are starting to fill this gap because neural networks are starting to be able to automatically identify by just reading tons of documents where things mean the same thing.

And so it's possible that the sources will go away in the next few years and be replaced by neural networks and a glossary. Don't forget a glossary because that's like the number one thing that people complain about. Like I don't know what that acronym means, can you please help me with the stupid acronym? So it's just a colossal area of acronyms you know so it’s helpful. Don't forget that these little things mean a lot to your employees, especially when you're on boarding somebody.

OK so let's get down to the final item which is really organizational structure which is knowledge graphs. And I'm a little bit schizophrenic about knowledge graphs. Part of me really loves them and wants to do them for everybody and part of me is very skeptical. Do you really need one? So thinking a little more skeptically: Do you really need to search for things related to things related to things? Sounds silly but people do, and it is actually a mission critical need to find exact relationships.

Are the relationships fairly well-defined? Do you have structured business data that can fill out at least some of the Knowledge Graphs? Or do you have to like build it from scratch from unstructured content? And you have a lot of money to spend? Is your knowledge graph an extension of an existing knowledge graph made to some by somebody else? Just be super careful about knowledge graphs. I think it's so easy to just say, hey let's do a knowledge graph and then you get into it, and 2 years later you're still building it and it hasn't actually helped your business in any substantive way. Now having said that, they're getting so much easier. The technology is getting easier; the editing methods is getting easier; the ways of building them are getting easier; and the flexibility of knowledge graphs is growing. So I think I might give you a kind of a different answer in a year, but right now I'd say approach them quite carefully. A Knowledge Graph is like an ontology but a little more flexible in that it's a bunch of links and nodes. Any node can be any entity and they're all put together. So it's not like you have customers in one table and organizational groups in another table and employees in another table. They all get mashed together into one structure and then you can just start linking things up. It's nice to have all these entities and even new ones you hadn't thought of, all just kind of existing together in a big graph. So if I have a customer and then I have the account manager for that customer, and then I have the projects for that customer, and then I have the people on those projects, and I have the technologies for those projects, and I have the PowerPoint presentations for those projects, all those can just be linked together in this big graph. And now you can do all sorts of cool things and ask questions, say who wrote the latest product? Show me the latest presentation for this customer. Who is managing this customer? Who's working on this customer? These sorts of things are great questions you can ask inside your knowledge graph. The real-life examples I've seen with crime - tell me everyone associated with this gun with the serial code. And so these are actual detectives trying to solve crimes or what parts of this chamber made of aluminum? Because you see aluminum particles on your materials. You're making something and you see aluminum particles so you want to say, well let's go through the diagrams in the charts and we make an whole knowledge graph of all the parts on the chamber. Who knows the CEO of this company so that you can identify and try and warm up the sales call. Or what US regulation was responsible for this compliance control? These are the sorts of real-life examples I've seen of actual knowledge graphs created to solve real business problems. There's a couple different types. I won't go into this; there's a kind of a religious difference I think between r.d.f. triples and label property graphs they both have sort of advantages and disadvantages. For the kind of more knowledge graph work I tend to prefer r.d.f. triples because they're a little more fluid and flexible rather than label property graphs which are a little more schema fixed. Having said that, property graphs can describe certain kinds of relationships that are impossible or very difficult to describe in r.d.f. triple. So you know there's no good answer and you see a lot of the knowledge work here have been centered around r.t.f. triples.

So there's seems to be a lot more lighter weight knowledge representation. r.d.f. triple seems to be where the trend is and the label property graphs tend to be for more specific examples, you know hard coded examples or narrow niche work. So here's an example from friend of a friend. This is an r.t.f triple example and so this is the node for James Wales who you know the cofounder we keep. And so there's his mailbox, his home page, his nickname, his depiction. We can see links. Now these are all links in fact, this is the name linked to a property which is his name, this is the mailbox linked to his mailbox, and then now these are links to other nodes linked to a person who's Angela Beesley.

And so that's a person linked through the friend of a friend. I presume once you have this, you can figure out how many degrees away from you know Kevin Bacon you are. So Knowledge Graph for me there's still a little too difficult to use. Some of the exceptions are if someone's already done the modeling for you. We have customers using a knowledge graph for pharmaceutical so that's a life sciences Knowledge Graph. A lot of customers are using that as that works extremely well. But otherwise I think you need to have a very compelling business case in order to spend the money on it.

So some closing remarks: the document, your knowledge structures, and then score to the final scorecard, thesaurus yes, controlled vocabulary no. Taxonomies and tags maybe, depending on the structure is it connected to, are they very clear? Will people actually use them? Knowledge graphs no, for now, but come back in a year and we'll see where that stands.

And thank you so much for your time I really appreciate it.