Catching the wave
Tools and Technology for Taxonomists
Taxonomy Bootcamp Washington DC
November 6, 2018
Three questions

Q1. What are taxonomists doing today?
Q2. How can new technology help?
Q3. What tools do you need to succeed?
What is it we taxonomists do?

categorize
organize
discover
Taxonomies are Knowledge Organization Systems.

When we do taxonomy we use industry standard data models to centralize and standardize the terminology used in our enterprise.

We define and unambiguously label enterprise terminology, and then we organize it into hierarchical and associative concept schemes, which we call taxonomies or ontologies.

These schemes help us to understand how concepts, people, places, products, processes and organizations all relate to one another.
Categorization enables an enterprise to retrieve, sort and rank content based on what it is about.

Concepts in taxonomies provide the metadata values for tagging documents and database records.

When we categorize content we use the semantics of our taxonomies plus contextualization rules to determine meaning and rank the relevance of content.

Taxonomy builds a bridge between the language people use to search and browse and the language found in documents.
The end-goal of what we do is we help people to retrieve more complete and accurate information and to discover latent knowledge.

Our work isn’t finished if we stop at categorization. We need to work with search teams and information architects to design and deliver semantic search and rich end-user discovery experiences.

including browsable navigation, faceted query refinement, and the ability to recommend related content.
Q1. What are taxonomists doing today?

Online survey – takes 10 minutes
link from pinned tweet at
https://twitter.com/DavidClarkeBlog

Participate in this survey during the conference and we will email you the revised results
Q1. How does your enterprise currently make use of taxonomies?

- Faceted query refinement: 66%
- Improve search accuracy: 66%
- Browsable navigation: 60%
- Lookup lists & glossaries: 44%
- Content classification: 42%
- Recommend related content: 36%
- Product Information Management: 28%
- Machine-reasoning: 18%
- Sentiment analysis: 6%
- Chatbots and conversational search: 6%

What we would expect... core taxonomy applications

Under-exploited... room for greater adoption

Participate today using link from pinned tweet at https://twitter.com/DavidClarkeBlog
Q2. Do your taxonomies contain general categories or specific concepts and names?

- Highly specific categories and subjects: 70%
- Broad categories and topics: 58%
- Individual named entities: 48%

Even mix of people doing broad-level categorization and highly specific subject indexing

Q3. How general or specific is your content tagging?

- Document tagging: 78%
- In-line tagging: 22%

Document retrieval more prevalent than page-level access

Q4. Where do you store your tagging data?

- Metadata in content systems: 70%
- Metadata in search engines: 48%
- Full text in-line document mark-up: 6%

Most taxonomy ends up as metadata in CMS / search

Participate today using link from pinned tweet at https://twitter.com/DavidClarkeBlog
Q5. How are your lexicons / taxonomies / thesauri / ontologies structured?

- Hierarchy: 86%
- Preferred and alternative labels: 50%
- Associative relationships: 44%
- Disambiguators: 38%
- Specialized ontology relationships: 22%

Under-exploited... room for greater adoption

What we would expect... core taxonomy

Q6. How public or private are your taxonomies?

- Within a company intranet: 52%
- Public-facing: 42%
- Shared with trusted partners: 20%
- Available for adoption and reuse: 14%
- Linked Open Data: 12%

Personal prediction... semantic web and automation will drive greater sharing of taxonomies even between commercial enterprises

Participate today using link from pinned tweet at https://twitter.com/DavidClarkeBlog
Q7. How do you create your taxonomies and/or use third-party taxonomies?

- Create our own in-house: 96%
- Reference third-party taxonomies: 26%
- Map to third-party taxonomies: 16%
- Public-domain taxonomies: 14%
- Linked Open Data taxonomies: 10%
- Tag using third-party taxonomies: 10%
- Commercial licenses: 4%
- Outsource to consultants: 4%

Most taxonomy still done within and specifically for the enterprise

Personal prediction... much greater adoption of LOD taxonomies

Q8. Which industry standards do you use for your taxonomies and/or for tagging?

- ANSI-NISO Z39.19 / ISO 25964: 44%
- Dublin Core: 32%
- W3C Linked Data: 20%
- W3C SKOS: 18%
- Schema.org: 14%
- W3C OWL: 10%
- W3C SKOS-XL: 2%

Miniscule adoption but this spec. means a lot to academic publishers and life-sciences

Enabler for machine reasoning but low adoption because of its complexity

Helps your metadata drive SEO

Participate today using link from pinned tweet at https://twitter.com/DavidClarkeBlog
Q9. What processes do you use to tag content?

- In-house professional indexers: 44%
- Self-tag content using taxonomies: 42%
- Self-tag using free-text keywords: 34%
- Tagging images, audio, and video: 30%
- Editable auto-categorization rules: 20%
- Human-supervised auto-categorization: 18%
- Fully automated categorization: 10%
- Training-sets and machine learning: 10%
- Outsourced content tagging: 6%

Human tagging is dominant.

Rules-based, human-supervised auto-categorization is more prevalent than fully automated or machine learning processes.

Participate today using link from pinned tweet at https://twitter.com/DavidClarkeBlog
Q10. Which emerging technologies do you think are already or will soon impact your enterprise?

- **AI / ML / Deep Learning**: 72%
- **Linked Data**: 58%
- **Big data and data science**: 54%
- **Semantic Web**: 52%
- **Graph databases**: 36%
- **Chatbots and conversational search**: 20%
- **Internet of Things (IoT)**: 18%
- **Blockchain**: 18%

No surprises given the hype around AI... but is the hype justified?

Linked Data identified as second highest impact on the enterprise... are you prepared?

Participate today using link from pinned tweet at [https://twitter.com/DavidClarkeBlog](https://twitter.com/DavidClarkeBlog)
Survey Insights

• Most (80%) tagging at the document level, fewer (20%) indexing to page/paragraph
• Over 40% of enterprise taxonomies are public facing
• What tech do people think will impact them most: AI followed by Linked Data
• But, currently there is low adoption (10%) of ontologies and graphs that enable AI
• Low adoption of schema.org, despite it being key to SEO for public-facing content
• Big surprise: most tagging still being done by humans
• Most common aspiration is for better quality auto-categorization

70% of enterprises are building traditional taxonomies and thesauri
18% are doing machine reasoning
Even mix doing broad-bucket categorization versus highly-granular subject indexing

Participate today using link from pinned tweet at https://twitter.com/DavidClarkeBlog
Q2. How can new technology help?

- AI
- Ontologies
- Linked Data

WHERE SHOULD WE FOCUS THIS YEAR? “BLOCKCHAIN” IT WILL CHANGE EVERYTHING.
EVERYBODY IS TALKING ABOUT IT.

THE POTENTIAL APPLICATIONS ARE ENDLESS. WE DON’T WANT TO BE LEFT BEHIND.
WHAT EXACTLY IS BLOCKCHAIN? ALSO, “ARTIFICIAL INTELLIGENCE”

Decentralized web Metadata management Records management Provenance & Governance IP rights management Supply chain management Linked Data

https://medium.com/@jimmysong/why-blockchain-is-hard-60416ea4c5c
How might the development of new technologies such as AI affect how we do our work as information professionals?

Primarily it should make some of the leg work easier. Being able to process a large amount of data in a shorter time than a human could, is going to be very helpful when it comes to the day-to-day work that we do. Machine learning and AI should also help us spot patterns in information that we may not otherwise notice - and it can help us simulate what the consequences of particular decisions might be.

However, **machine learning and AI won’t be a silver bullet** - we’ve already seen examples of algorithms being applied with unfortunate consequences, or machine learning classifying images in a way that is problematic when we consider culture, race and politics. So, we’ll need to be aware of the limits of what is possible, manage expectations, and take on new responsibilities for helping machines understand our world, our biases, and our morality.


What's the most exciting change you've seen in the industry in the last few years?
The deployment of knowledge graphs and non-visual interfaces have really brought the value of structured data to the fore. Voice and conversational interfaces are obviously very in vogue at the moment, and I think an under-appreciated aspect of these has to be the importance of structured data in powering these. I think more and more people are seeing the benefits of structured data in things like Google’s Knowledge Graph, Amazon’s ability to tell you which actors are on screen during a show you’re watching, and of course the perennial joy of doing a deep dive on Wikipedia.

Whilst ‘big data’ and ‘machine learning’ might be hogging the limelight at the moment, I think that the work of taxonomists and those who architect and develop structured data is quietly, gradually, revolutionising the kinds of things we do with computers and the Internet.
Hi Dave,

Smarter AI requires even smarter data, but how does that data stay sharp in a rapidly changing industry? Sometimes it takes a human touch to bridge the gap, and curated crowds are the perfect source for that bit of human tuning.

“Behind all this AI, humans still need to touch the data. We still need human power to teach computers the subtleties of identifying data, language, dialects, and so on.”

— Kerri Reynolds, SVP Human Resources & Crowdsourcing, Appen

“A fully automated generation of ontologies from text corpora is not possible and won’t be possible in the next couple of decades. This is the same kind of AI promise that has failed many times before... Generation and maintenance of taxonomies and ontologies will always remain to the realm of human beings and their knowledge of the world.”

— Andreas Blumauer – Pool Party

“...we’ve increasingly come across organizations that have been promised Artificial Intelligence (AI) capabilities, but have not realized them... The message we consistently hear, however, is that these [AI] tools haven’t lived up to the promise. Though the demos are impressive, the reality is deflating.”

— Zach Wahl – Enterprise Knowledge

“Growing interest in AI and machine learning and its impact on indexing and tagging. I feel there are limits in automatic technology and taxonomy creation.”

— Heather Hedden – Author The Accidental Taxonomist
There are many different types of AI, and different types are more or less suited to different business applications, specifically some are just not relevant to taxonomy and semantics...

...learn about different use-cases with this interactive online guide from McKinsey

AI is very relevant to what we do, but it will not replace the need for human curated taxonomies or ontologies.

On the contrary, it is taxonomies and ontologies that will empower AI with the semantics and logic to improve search, categorization and machine reasoning.

Make sure management in your organization understand this.
Do you need ontologies...
...depends on the type of problem you are trying to solve an analogy...

Taxonomies have a singular top-down way of organizing information

Now try modelling the Washington Metro as a taxonomy... it won’t work because it is a graph

...great for a classification or guided browse experiences

...ontologies are graphs, they support many points of entry and many alternative pathways
The boundary between Taxonomy and Ontology can be a confusing grey area...

...distinguishing Property from Value Vocabularies may help

...we take an holistic approach and use ontology to design and build taxonomy
Linked **Open** Data: a huge and growing set of re-useable public domain ontologies and taxonomies

Linked **Enterprise** Data: enterprises can benefit from the data model even when their data is not open
Ontologies and Linked Data are highly relevant and offer many benefits including:

+ **Build smarter search and discovery applications** by leveraging the logical dependencies defined by ontologies.

+ **Reduce costs and speed up project deliverables** by reusing a vast and rapidly growing library of public domain property vocabularies (ontologies) and value vocabularies (taxonomies).

+ **Simplify systems integrations** by adopting open industry standard data models and portable data interchange formats.
Q3. What tools do you need to succeed?

Designing and building standards compliant taxonomies and ontologies and developing good categorization rules isn’t easy…

…good software tools will simplify the complexity.

**organize**

Taxonomy, Ontology and Linked Data Management Systems

**categorize**

Text Analytics, Auto-Categorization & Human Tagging Systems

**discover**

Semantic Search, Recommenders, Visualization & Chatbots
What to look for in Taxonomy Management Systems

- Drag-and-drop editability
- Collaboration & governance
- Automated management reports

See also Synaptica’s Top 100 Features Checklist at https://www.synaptica.com/resources/

... and the ability to switch seamlessly between taxonomy editing and categorization rule editing
What to look for in Ontology & Linked Data Management Systems

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>An extensible library of public domain ontologies</td>
<td>Design your own taxonomy schemes using plug-and-play ontologies</td>
</tr>
<tr>
<td>Easy-to-use UI/X to simplify the complexities of...</td>
<td>... generating standards-compliant RDF triples and knowledge graphs.</td>
</tr>
</tbody>
</table>
What to look for from Categorization & Text Analytics Tools

Categorization rules that are transparent easy to edit without having to learn esoteric syntax...

... a no-black-box principle allowing users to understand how rules work and quickly refine them.

Seamless integration between taxonomy management and categorization management workflows...

... a no-silo approach to reduce complexity and improve productivity.
What to look for from Search and Discovery Tools

- **NLP to handle natural language queries.**
- **Semantic search to improve precision and recommend related content.**
- **Taxonomy-driven IA including facetted navigation and query refinement.**
Software tools exist to help taxonomists and indexers be more effective and productive... **they must simplify the complex.**

They must **use open industry standards** for data portability and interoperability.

They should help taxonomists to become experts **without having to learn esoteric code.**

**Search needs to take advantage of the semantics of taxonomy** and ontologies to deliver smarter applications and a richer knowledge discovery experience.