

# Leveraging Collective Knowledge

Dr. Henry Kon

Intellisophic

15 Maple Ave

Paoli, PA 19301

hkon@intellisophic.com

Michael Hoey

Intellisophic

15 Maple Ave

Paoli, PA 19301

mhoey@intellisophic.com

## ABSTRACT

As more organizations begin to deploy taxonomies for categorization and faceted search, the cost of producing these knowledge models is becoming the largest expense on a project. At a cost of 200 – 300 dollars per topic, manually developing subject area taxonomies does not scale for any but the smallest of projects. This paper will discuss an approach called Orthogonal Corpus Indexing (OCI). OCI leverages existing published knowledge in the subject area of the taxonomy model. This knowledge is algorithmically mapped into multiple taxonomies via the OCI algorithm. The resulting taxonomy costs are 1/ 100th of the cost of manual methods and are created with embedded rule sets for categorization engines. This paper will discuss the theory of OCI, its practical use as well as examples of knowledge management techniques that are possible when taxonomies are large, detailed and inexpensive.

## Categories and Subject Descriptors

I.5 PATTERN RECOGNITION, I.5.5 Implementation, H.3.3 Information Search and Retrieval

## General Terms

Algorithms

## Keywords

Taxonomy, Library, Classification, Categorization, Tagging

## Introduction

### 1.1 TAXONOMY FOR INFORMATION ORGANIZATION

- *Is it feasible to assemble taxonomic data covering all human knowledge?*
- *What is an appropriate concept representation?*
- *Should concepts be assembled a priori? Or is a dynamic model possible?*
- *How will quality be achieved and maintenance governed?*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'05, October 31-November 5, 2005, Bremen, Germany.

Copyright 2005 ACM 1-59593-140-6/05/0010...\$5.00.

Evidence suggests that the loaded cost of developing a taxonomy may be as high as two hundred dollars per node, which seems reasonable when considering the involvement of IT staff, corporate librarians, departmental publishers, commercial information providers, and international standards bodies. The questions above are informed in various ways by *orthogonal corpus indexing* (OCI), the methodology and technology discussed herein for automated development of concept taxonomies from existing reference text corpora. OCI has resulted in a large scale concept catalog and taxonomic library, leveraging existing works for authorship, as a way to exponentially reduce taxonomy development costs.

### 1.2 TAXONOMIES AND CONCEPT INDEXES

Semantically rich knowledge representations do exist, and several are emerging for semantic interoperability and web services such as W3C's OWL<sup>1</sup> and DARPA's DAML<sup>2</sup>. There is, however, no global ontology of concepts to populate these schematic representations. High degrees of up front design and pre-coordination are required to affix concepts to data. And while taxonomies are intended to enable conceptual access to information - the effort, expertise, and risk involved in manually generating taxonomies results in a lack of rich, deep, and extensible taxonomic data to characterize the conceptual contents of underlying text. Probabilistic information retrieval<sup>3</sup> thus requires up front articulation of each concept space.

Human knowledge naturally covers many domains, ranging from general interest such as entertainment and foods to the sub-sub-topics of specific legal or scientific sub-specialties. OCI is distinctive as a taxonomy generation method in that it leverages structured reference texts for development and extension of an integrated topic space enabling deep yet accurate taxonomies on a scale never before achievable. In a matter of hours or even minutes a source corpus can be converted into taxonomic data through automated methods, with semi-automated controls for quality enhancement as desired. In general, taxonomy development is akin to data modeling in that a conceptual model is developed to cover the domain of interest. Data modeling is done a priori to meet anticipated application requirements. During taxonomy design for a given industry or segment, if a subject area is not already in our library, we identify relevant and accessible publications to cover the domain of interest.

An *orthogonal reference text* corpus for OCI input can be any body of text that is organized by concept or is otherwise topically

prearranged with topic delineations such as chapter headings or a table of contents. Sources used include encyclopedia, reference texts, handbooks and training manuals, gazetteers, on-line directories, or any number of content types and sources satisfying the base requirements. The source reference publications studied usually are alphabetical or have a hierarchical topic graph with distinct and authoritative text for each topic. After structure recognition of the source corpus, the text portions from that corpus are processed to derive concepts in the resulting taxonomy.

The source corpus provides three inputs:

- a) taxonomic structure (typically largely hierarchical) via an outline structure or other topic delimiting mechanism,
- b) a title term as a label for each node, and
- c) text describing the topics or concepts from which is derived a signature vector containing facets –a facet may be a word, phrase, entity identifier, concept identifier, or other domain-specific information derived from the source text.

Figure 1 below is an example taxonomy and concept derived from *Clarke's Analysis of Drugs and Poisons*<sup>4</sup>. In print form, this source corpus is a 2100 page reference source on drugs, poisons, and related substances. It is aimed at scientists attempting to determine a drug or poison in a pharmaceutical product, in a sample of tissue or body fluid, from a living patient, or in post-mortem material. *Volume 1* (shown below) has 32 chapters concerning analytical procedures in forensic toxicology. *Volume 2* has more than 1750 substance monographs detailing physical properties, analytical methods, pharmacokinetic data, and toxicity data. The graphic shows the concept of 'Pharmacokinetics' within the reference corpus. The taxonomy below is opened up to this node four levels down in the hierarchy of *Volume 1*. On the left is the concept of 'Pharmacokinetics' in the context of the topic hierarchy around it, showing a navigational path to it.

**Figure 1: Example taxonomy and 'Pharmacokinetics' node with term signature facets**

The concept shown is given context by the path location:

```
> 1. Clarke's Analysis of Drugs and Poisons
    > 1.6. Postmortem Toxicology
        > 1.6.2 Analytical Toxicology
            > 1.6.2.5 Pharmacokinetics
```

On the right is shown the set of facets derived by OCI from the corpus source on this topic to describe it. The source text for this concept contains text from which OCI text processing derives

discriminating facets starting with the term "postmortem cases". Each facet is weighted according to the relative importance of that facet in that topic's source text in the corpus. Signature facets are the words and phrases, hyphenated terms, entities, or other possibly domain-specific elements such as chemical formulas and concepts derived from the text. Facet derivation is achieved using statistical and linguistic processing including phrase generation, parsing rules, lexical filters, and other tools such as entity extraction and categorization

### 1.3 CONCEPT COUNT AND CONCEPT RICHNESS

This rich multifaceted representation of an OCI concept is atypical relative to what industry refers to as a node or concept in a taxonomy. Figure 2 below illustrates the position of OCI relative to alternative approaches to taxonomy development – both automated and manual. The left side of the table concerns the richness of a concept: is a concept represented as a single term, e.g., an undifferentiated string, or is a concept multifaceted. Auto-categorization and other concept operations require a rich concept representation in order to go beyond keywords as a proxy for concept matching. The horizontal on the table concerns the number of concepts achievable - low being on the order of tens of thousands or less, while high is above tens of thousands. The higher the number of concepts, the more subject areas and nuances within them are covered, and thus the deeper the taxonomy can be developed into specialized areas.

No technique besides OCI offers the accuracy and richness of authoritative concepts as well as the depth and breadth of accurate taxonomic hierarchies. The fingerprint of each concept is captured in the signature facets and reflects the authorship and expertise of a subject matter expert (the corpus author). This elaborate articulation of concepts across a large concept space enables:

- enhanced precision and recall in auto-categorization and document filtering
- richer search keys and concept interconnections
- cross-domain research and discovery of intellectual property
- deep tagging and summarization of content with concepts
- cross-referencing of documents, concepts, and people.

	<b>Low Node Count</b> – not deep or specific in specialized areas < 10,000	<b>High Node Count</b> > 10,000
<b>Low Concept Richness</b> concept is represented as a String	<u>Manually developed</u> thesauri and taxonomies for narrow domains or specific applications	<u>Industry thesauri</u> (several exist with tens of thousands of terms) <u>Library of Congress, Dewey Decimal</u> <u>Yahoo!, Open Directory</u>
<b>High Concept Richness</b> Multidimensional/multifaceted concept representation	<u>Clustering</u> (e.g., <u>latent semantic indexing / SVD</u> ) (inaccurate/unlabeled concepts, methods require redundant sources, irregular taxonomies, limited application) <u>Training sets</u> (must manually define hierarchy and collect documents per node)	<u>OCI</u> (millions of multifaceted and accurate concepts, authoritatively organized and described, supports deep and meaningful auto-categorization, extensible concept space)

**Figure 2: Rich vs. Deep Taxonomy Creation Methods**

## 1.4 TAXONOMY PUBLISHING

OCI processing begins with a corpus source of reference information. Leveraging prior corpus creation by authors, editors, and publishers reduces the direct cost

of developing a taxonomy concept by several orders of magnitude while enhancing taxonomy quality in several key dimensions. By selecting corpora from commercially published sources, and complemented by various other sources, Intellisophic has assembled a rich taxonomic library with millions of concepts and facets.

Consider the breadth of reference material available as taxonomy sources. There are several thousand English language encyclopedia published, and there are tens of thousands of professional and technical reference books. Encyclopedia topic areas currently populated range from general knowledge such as *WorldBook Encyclopedia*<sup>5</sup> with twenty-seven thousand topics, to technical and professional specialties such as the *Encyclopedia of Chemical Technology*<sup>6</sup> with its forty thousand topics in a twenty-seven volume set and on to a variety of more specialized areas. OCI has also been applied to corpora of various genres such as patents, legal acts, government reports, on-line information directories, and internal training manuals.

When a topic area is not covered or is not sufficiently elaborated and taxonomy extension is desired, this is achieved through access to new corpora. OCI enables rapid creation of high quality taxonomies on demand.

These examples illustrate the massive collective knowledge resource that existing works – both formally and informally published - offer for taxonomy development and extension. This off-the-shelf labor saving model allows taxonomies to be extended as application requirements change and as knowledge grows into virtually any area of human activity. In contrast, the effort required for manual pre-coordinated taxonomies forces organizations to choose one concept and term at a time, in the

context of a given application - by librarians, subject area experts, and editors who must agree on terminology. This process is tedious, with results that are often rigid, of low coverage, or otherwise of low quality.

## 1.5 DEVELOPMENT METHODS AND TOPIC COVERAGE

*“Lincoln, the Man, the Car, the Tunnel”  
was a play about romance”*

Keywords are unreliable as indicators of conceptual content as the above example illustrates. This is due to the all-or-nothing keyword approach in which a document either does or does not have a keyword. The mental construct of a “concept”, however, rarely corresponds to a single keyword or to a Boolean combination of keywords. A biochemist looking for information on a particular chemical process might search by inventor name, an input compound, a regulatory group, an acronym or common name, or any of a number of possible and probable handles into the conceptual space. Synonyms and alternative wordings further compound the problem. A taxonomy provides a basis for conceptual access and sharing of information as well as enabling mining of textual data for business intelligence, whether that text is in a document repository, newspapers, emails, or embedded in the text fields of a database. A concept within a taxonomy sets context and must provide for a rich multifaceted representation to overcome the naïve model of keyword matching.

## 1.6 ENUMERATED AND PRE-COORDINATED TAXONOMIES

Although a database installation may have hundreds of millions of rows, the set of metadata descriptors required for data integration and mining, such as authoritative terms or controlled vocabularies, is relatively bounded. This is because most database applications are targeted to a well-defined domain of discourse. Thus, metadata models are usually developed by a centralized standards group for

structured data. Categorization schemes for unstructured text such as the Library of Congress Classification (LCC)<sup>9</sup> and Dewey Decimal Classification (DDC)<sup>10</sup> are considered *enumerative*, meaning that a notation is assumed to exist to categorize any given publication. The LCC came into being in the early twentieth century and is used in most academic library collections. Indexing elements include: subject, author, date, geography, and genre. Subject is the primary sort key, and the job of the human categorizer is to select a category for the particular information package. In the LCC system, the first letter of a call number refers to the general subject area and the second letter refers to a subsection within the general subject category. For example, a book on the History of Modern France can be categorized as first level World History and second level France.

Such notations are called *pre-coordinated* because there is prior agreement on the structure and subjects of interest. The centralized process to define and maintain a categorization scheme involves governance and stewardship by committee; a distributed large-scale standards approach. The major drawback to this model is that there is a rigid network of paths leading to rigidly grouped items. The job of the categorizer is to “categorize each book into pre-existing pigeon holes”<sup>11</sup>. While useful for placing books on a shelf, and thus for high-level categorization, schemes such as the DDC or the LCC are not intended to detect and represent the full spectrum of subjects in each information package.

So while pre-coordinated and enumerated taxonomies, thesauri, and lexicons have a place for book identification in library science, their role in automated concept recognition is less clear. Often there is little prior knowledge about the set of topics contained in a document or document pool, and no opportunity for people to read and categorize manually, particularly in a multifaceted context. The goal may be to discover emergent concepts in email, text notes, or a text message stream. The OCI Model for content indexing is not to bound, control, or pre-specify a restricted set of terms and concepts - but rather to allow the space growth in a more organic sense, to use systems to impose organizing structure and for meaning extraction. The resulting rich and deep taxonomic data are high-dimensional indexing structures for connecting to information conceptually.

## 1.7 MULTIFACETED CONCEPTS

Ranganathan<sup>12</sup> described the Dewey Decimal system as faulty in its underlying principle of listing all possible subjects, assigning to each a predetermined categorization number, and subsequently fitting everything into existing buckets. Ranganathan saw that human pursuits were growing quickly. New areas of knowledge were being discovered, and new ways to combine existing subjects were emerging as well. Any categorization scheme that attempted to enumerate a finite number of subjects, without proper capability for expansion into new knowledge, could never meet the needs of the future<sup>13</sup>. Ranganathan's seminal work in library and information science formalized *multifaceted concepts* and the *Colon Classification Model* - a model for defining concepts as combinations of other concepts.

For example, the Colon system would represent a book about “research in the cure of tuberculosis of lungs by x-ray conducted in India in 1950” with a call number as follows:

L,45;421:6;253:f.44'N5

The notations represent:  
Medicine,Lungs;Tuberculosis:Treatment;X-ray:Research.India'1950<sup>14</sup>. This is in contrast to the Dewey Decimal and Library of Congress systems which require new concepts to be enumerated individually. The Colon system was a major theoretical development for categorization theory. It acknowledges the various ways that concept facets relate to one another and showed that concepts can be effectively combined to derive new concepts. The rich concepts derived by OCI are multifaceted, as each corpus article is processed to extract the essential facets of that topic. Often twenty or more distinct facets of an article are derived as signature elements.

## 1.8 WHY MILLIONS OF CONCEPTS? “SPECTRAL CONTENT ANALYSIS”

A node within a library categorization scheme is not intended to represent the spectrum of concepts contained in a publication, but rather to reflect the general subject of that publication. A book can be put only one place on the shelves, even if it contains hundreds or thousands of underlying concepts. So how many concepts might be needed? A concept, for OCI purposes, can be interpreted as a domain of discourse and may include either base or aggregate concepts such as:

- a particular event in history
- logistical issues in the transportation and storage of citrus fruits
- taxation on oil production in Europe
- health effects of inhalation of depleted uranium
- a particular famous person
- the general concept of a famous person or of an athlete

Within source corpus text, ideas are represented as strings in a natural language: paragraphs, sections, and section headers are often the only structure imposed on otherwise ambiguous words, phrases and sentences as rendered by the corpus article author. Eventually each facet will be derived via OCI processing from corpus article text. Abstractly, the facets of a concept that OCI extracts are the various entities, attributes, and relationships that would be used to talk about that concept. Facets are not tightly bounded, as they are intended to provide an interconnect model to relate different concepts to one another, and to search and navigate among concepts and documents.

To estimate the number of concepts or facets in the world at large, or as relevant to some particular domain, is only a theoretical exercise. To a child, a rock is something to be thrown, while to a geologist it represents something much more. Nevertheless, many concept labels require multiple terms, for example: *health effects of inhalation of depleted uranium* uses multiple significant terms. Concepts aggregate in a multiplicative way to become more particular or aggregated. Additive as well as subtractive models of concept definition are relevant, for example a concept specification may indicate that someone is interested in *anthrax* the *bacteria* and *industrial hygiene*, but not in the context of the concepts of *cattle* or *livestock vaccination*.

The Oxford English Dictionary<sup>15</sup> contains approximately 500,000 terms, not including another estimated 500,000 scientific terms<sup>16</sup>. Clearly not all combinations of terms and concepts would result in a meaningful aggregate concept. Regardless, the combinatorial aggregations of words into concepts are huge even for highly restricted combinations. WorldBook Encyclopedia has 27,000 topics. The *Encyclopedia of Chemical Technology* has 40,000. *Medical Subject Headings* by the National Library of Medicine<sup>17</sup> has 50,000 terms. The Encyclopedia of Modern Asia<sup>18</sup> has 6,000 topics. The Gale Business Thesaurus<sup>19</sup> has 15,000 terms. Virtually any specialty area, well articulated, can quickly climb into thousands or tens of thousands of concepts. The goal of a concept catalog should be to preserve and represent as much information and detail about a domain as possible, and only through a fine-grained concept index is this possible. The set of concepts relevant across organizations, across document collections, and across applications is clearly too large to manually enumerate into multifaceted representations for automated processing.

Thesauri – often called taxonomies in industry vernacular – are developed for authoritative term sets and for search enhancement. Their entries are what can be called *atomic concepts* – concepts which correspond tightly to the terms of the language. But, as discussed, many concepts are synthesized from multiple underlying or related concepts and thus can never appear in a thesaurus. Comprehensive authorship about virtually any field is not simply itemizing base and related terms; it is the discussing of areas of knowledge in various ways: the components of a system, the steps in a process, the disciplines underlying a scientific challenge. Consider these example topics or concepts from reference corpora Intellisophic has processed:

- Detecting Foodborne Pathogens
- Misinformation On-line
- Anti-fungals for Dermatological Use
- Surveillance of Personnel for Laboratory-Associated Rickettsial Infections
- History of the Federal Bureau of Investigation

Concept taxonomies must be considered dynamic structures which can grow as needed as knowledge evolves and, in a sense, are not even relevant to the thesaurus model. OCI processing has resulted in a rich taxonomy library that has millions of such concepts and underlying facets, articulating the subtlety and interconnectedness among concepts. Intellisophic continues to grow the library into new and more detailed topic spaces.

## 1.9 OCI TAXONOMY COVERAGE

The breadth and depth of the OCI taxonomy library are driven by two factors – content availability and application demand. The current strategy is to cover general knowledge with multiple general encyclopedias and to build out increasingly targeted taxonomies into more specialized areas. Source corpora range in size from several hundred to as many as 100,000 topics.

Sources have included a variety of genres including encyclopedia, dictionaries, controlled thesauri with definitions, handbooks and manuals, professional and reference textbooks, directories, and patent abstracts.

This content is evolving with a growing set of source materials to meet the demands of new applications, and over time with a growing base of human knowledge. As for quality – if inadequate corpus content is fed to the system (e.g., insufficient text to describe a topic) then quality will suffer, though taxonomy structure and labels may still be robust. The algorithms employed on the content processed to date have shown to be an extremely effective combination. While topics need not be mutually exclusive, it is believed that the taxonomy library assembled thus far is the broadest, deepest and richest in the world.

With such a large library, there is some topic aliasing. This is seen as a benefit in knowledge mining applications. Let's take a real world example: I have two topics titled Gastrointestinal Cancer, one in a taxonomy on diseases, and another on taxonomy specifically covering oncology. If I am looking at diseases in general, I may only require the depth that a single topic on the condition gives me. If I am interested in cancer, then I need more detail and more topical coverage that I would get from the title specifically on the subject. The oncology taxonomy may have 30-40 topics on this area of cancer. If I want to go deeper still, I can select taxonomy titles that may have over 1000 topics on a specific type of cancer. This capability of OCI, is one that makes it powerful for knowledge mining applications. The taxonomic axiom of one topic only in one place owes its roots to the placement of books on shelves where the physical world limited how we found and used information. This can in fact be a limiting approach in knowledge mining applications where placement is just a row in a database.

## 1.10 TAXONOMY AND CONCEPT REPRESENTATION

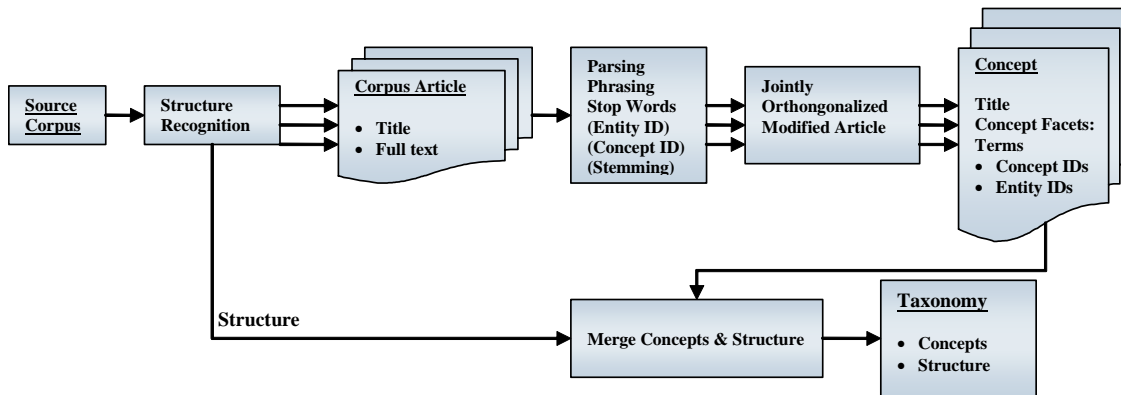
An OCI concept is a node within a largely hierarchical graph structure of topical parent-child relationships. Other relationships are also possible such as geographic containment, experts within a discipline, etc. These are determined by the semantics of the source corpus and do not necessarily alter the processing performed. A node corresponds to a concept and contains a title, a parent pointer, a signature vector of *concept facets* and their weights, and optionally children and link references. While the term *hierarchy* is used to refer to Intellisophic's graph model, secondary link types are also supported for constructing a new taxonomy from existing ones as well as for typed links besides parent-child such as "see also" links.

A *facet* is an aspect of a concept, i.e. a construct used by the corpus author in discussing that concept. At a modeling level, three facet types are relevant:

- 1) A term, which is a word or phrase lifted from the text with no identified relation to an outside construct (an undifferentiated string).
- 2) An entity, which is a reference to an instance in an entity extraction and identification system (e.g., ID's of specific people, places, books, organizations, dates, etc.).
- 3) A concept, which is to say that the corpus article itself contained a specific pre-existing concept.

The faceted representation of a concept defines each concept as a vector in a multidimensional space of dimension  $n$ , where  $n$  is the number of all terms, concepts and entities defined.

## 1.11 CORPUS PROCESSING



OCI processing looks at a corpus to determine which elements of each text article belongs in the final facet set. As shown in Figure 3, a source corpus is first processed for structure recognition, which may include the processing of a table of contents or the identification of section headers. The text from each section is parsed and processed to determine the set of facets for possible assignment to a given concept. The text processing to identify facets includes parsing, phrasing, and elimination of stop words.

Matrix processing allows for rapid handling of the large multidimensional space of facets and weights to determine which facets are to be assigned to which concepts. The orthogonalization step identifies a unique signature for each concept. The set of facets which is associated strongly with a corpus article is assigned to that concept and thus defines the characteristic signature or fingerprint for the derived concept.

Results can be stored in a relational database for easy access and connection to other semi-structured and structured data elements within the enterprise. The taxonomic data that OCI generates are delivered in any of several external formats including the XTM Topic Map standard<sup>22</sup> and the W3C Web Ontology Language.

## 1.12 APPLYING DEEP TAXONOMIES WITH RICH CONCEPTS

A data model represents the structure and vocabulary of data elements for a given database application. Consider ISBN numbers in the publishing world. All applications using this model must commit to the ISBN as a shared representation, without which bookstores could not communicate about books. Of course, ISBN is a narrow domain. To build applications around constructs that have not been pre-established, and may grow rapidly, requires an extended vocabulary and modeling ability. The area of *text mining* concerns the extraction of implicit, unknown (e.g., to the corpus authors), and useful information from large amounts of textual data. Applications enabled by deep taxonomies with rich concepts deal with large bodies of text, concern new or hidden information (e.g., discovery), and involve a domain of concepts that can only be specified for application design at an aggregate subject-area level.

## 1.13 CATEGORIZATION, SEARCH AND NAVIGATION

The problem of categorizing a document against a set of concepts is to either:

- 1) determine the single high level concept most relevant to, or representative of, the document (e.g., the library bookshelf / categorization problem),
- 2) identify a limited set of concepts that are most relevant (e.g., beyond some scoring threshold to capture essential elements of the document), or
- 3) list all concepts contained within the document (yielding an inventory of the document for auditing or search purposes).

Categorization on the basis of nodes represented as a single undifferentiated string as the concept representation, or with a shallow topic hierarchy and non-specific concepts, would suffer similar problems to keyword search; namely that scoring for relevance would be unstable, unreliable, or simply uninteresting for their generality. False positives and false negatives result from synonyms, homonyms and hyponyms (a word that is more specific than a given word). Categorization against a multifaceted concept allows scoring to be multidimensional and therefore to more closely reflect subtle human judgment about conceptual relevance.

Take into account again the concept of 'pharmacokinetics' from Figure 1. The top-level signature facets are the terms 'postmortem cases', 'calculations', 'concentrations of a drug', 'adipose', 'toxicologists', and the concept label 'pharmacokinetic' itself. In this manifold representation, much more than simply the concept label is surfaced. The author's perspective, as to what is essential to the topic and thus *what the topic means*, comes through in the OCI concept representation. It can be seen from the signature terms that this article was not focusing on other aspects of the topic such as *mechanisms of drug diffusion* or *detection of drugs within a pharmaceutical*. Rather, this article describes

'pharmacokinetics' in terms of concentrations of drugs in body tissue in postmortem cases.

Considering another example, the signature terms in one context for the concept of 'sterilization':

*steam, sterilize, germs, viruses, surgical, killing, medicines, bacteriology, virus-like*

And another context for 'sterilization' (from a different taxonomy) has the terms:

*procedure, surgical, reversal, tying, tubes, anatomical, blocking, sterilized, castration, spaying, contraception, ovaries, vasectomy*

These divergent signature sets reflect the fact that the term 'sterilization' could refer to the removal or destruction of microorganisms, or it could refer to the act of making an organism infertile. Clearly 'sterilization' alone is ambiguous. The distinct signature terms make very clear the differentiation between these two uses of the term.

Relevance scoring of documents against a multifaceted concept is a mathematically based alternative to complex Boolean queries and has been shown to increase both precision and recall significantly. Government labs testing based on the TREC Q&A track material<sup>23</sup> have indicated significant improvements relative to competing content and categorization models. Based on a multidimensional relevance scoring algorithm, not only would the term sterilization be used, but a joint vector cosine score and signature coverage score can be applied. Multidimensional scoring against a concept signature is a robust way around keyword indexing with its complex and contrived Boolean combinations of terms and query design issues.

This same multifaceted representation also aids in search and navigation, in that multiple paths and interconnections can be identified. An extended keyword search on signature terms may aid in a discovery process for a researcher investigating toxicology, drugs, forensics, etc. Having the concept term 'pharmacokinetics' placed in the context of

*Clarke's Analysis of Drugs and Poisons > Postmortem Toxicology > Analytical Toxicology > Pharmacokinetics*

enables fast disambiguation of faceted terms and concepts, as well as introducing further navigation opportunities. Similarly, 'sterilization', is easily disambiguated or discovered based on navigational paths.

*General Knowledge > Life Science > Bacteriology > Sterilization*

vs.

*Health > Reproductive Health > Birth Control > Surgical > Sterilization*

## 1.14 EXAMPLE USE - NARRATIVES

Another take on multifaceted concepts is the *narrative*. A narrative is a document selected by a user to represent some area of knowledge or discourse, i.e. a subject of interest. For a

researcher at a pharmaceutical company, this may be the description of a new drug that is under development. By categorizing such a narrative document against a set of taxonomies (say covering several hundred thousand topics in medicine, diseases, adverse events, mechanisms of action, and chemistry.), the underlying set of concepts contained in the narrative immediately become clear. The concept decomposition of this narrative provides a review of the conceptual contents of the document. The resulting collection of concepts is a structured data view of what is contained in the document. The structured data set, in the form of a concept inventory detected within the narrative, is now a basis for discovering and identifying more information about that same narrative constituting a conceptual "more like this" feature. The difference is that we can now use structured data tools to evaluate the similarity of the narrative to other documents that are being mined. We are also basing the similarity, not on word patterns, but on concept patterns. This lift gives us significantly more precision, because it is based on the core knowledge contained in the taxonomies. From an IT perspective, we also have a much simpler computational problem, because we are dealing with structured data, as opposed to unstructured. From a functionality perspective we have two enhancements. One, is that the approach of using narratives along with subject matter taxonomies yields answers to problems that cannot be approached with traditional full text engines. The other is that to the end user, the taxonomies have become a knowledge infrastructure. An end user is not using the taxonomies for navigation, but rather to embed the core knowledge of the subject area into the system. As a user, they do not need to be familiar with the taxonomies, or even see them. They need only be able to describe what they are looking for (a drug similar to the one they are studying) and the computer will compare this with the subject area taxonomies and mine the data in the structured data domain.

## 1.15 CONCLUSION

This paper has introduced Orthogonal Corpus Indexing as a model for efficient production of high quality taxonomies, and described several applications for the kinds of taxonomies that OCI produces. Existing vetted sources impose structure and identify concept labels and facets resulting in rich, deep and meaningful taxonomies achievable through automation. Spurious or unlabeled "concepts" that may arise from automated approaches such as clustering are not an issue in this approach. The quality of the resulting taxonomies is high for breadth, depth, accuracy, balance, consistency, extensibility, and interconnection.

## REFERENCES

- <sup>1</sup> Web Ontology Language (OWL) <http://www.w3.org/TR/owl-ref/>
- <sup>2</sup> DARPA Agent Markup Language (DAML) <http://www.daml.org/>
- <sup>3</sup> Maron, M. 1961. Automatic indexing: an experimental inquiry. *Journal of the Association for Computing Machinery* 8, 3, 404–417
- <sup>4</sup> *Clarke's Analysis of Drugs and Poisons*, The Pharmaceutical Press, 2003
- <sup>5</sup> WorldBook Encyclopedia [www.worldbook.com](http://www.worldbook.com)
- <sup>6</sup> Encyclopedia of Chemical Technology, 27 Volume Set, John Wiley, q2004
- <sup>9</sup> Library of Congress Classification System
- <sup>10</sup> Dewey Decimal Classification System
- <sup>11</sup> Vickery, B. C. Faceted Classification Schemes. *Rutgers Series on Systems for the Intellectual Organization of Information*, ed. Susan Artandi, v. 5. New Brunswick, N.J.: Rutgers University Press, 1966
- <sup>12</sup> Ranganathan, S. R. *Elements of library classification*. Bombay: Asia Publishing House.1962
- <sup>13</sup> Satija, Mohinder Partap. *Colon Classification, 7th edition : A Practical Introduction*. New Delhi : p.2, Ess Ess Publications, 1989
- <sup>14</sup> Glassel, A. <http://scout.wisc.edu/Projects/PastProjects/toolkit/enduser/archive/1998/euc-9803.html>
- <sup>15</sup> *Oxford English Dictionary*, Oxford University Press
- <sup>16</sup> McCrum, R., Cran, W., & MacNeil, R. *The Story of English*, New York: Penguin, 1992: 1
- <sup>17</sup> *Medical Subject Headings* by the National Library of Medicine <http://www.nlm.nih.gov/mesh/meshhome.html>
- <sup>18</sup> *Encyclopedia of Modern Asia*, Charles Scribner's Sons, 2002
- <sup>19</sup> *Gale Business Thesaurus* <http://www.taxonomywarehouse.com/vocabdetails.asp?vVocID=3>
- <sup>22</sup> XTM Topic Map 1.0 Standard <http://www.topicmaps.org/xtm/>
- <sup>23</sup> Text Retrieval Conference, Q&A track, <http://trec.nist.gov/data/qa.html>