‘NATURAL CLASSES’ IN THE KNOWLEDGE WEB

Michael K. Bergman[1], Coralville, Iowa USA
July 13, 2015
AI3:::Adaptive Information blog

We have recently talked much of the use of knowledge bases in areas such as artificial intelligence and knowledge supervision. The idea is to leverage the knowledge codified in these knowledge bases, Wikipedia being the most prominent exemplar, to guide feature selection or the creation of positive and negative training sets to be used by machine learners.

The pivotal piece of information that enables knowledge bases to perform in this way is a coherent knowledge graph of concepts and entity types. As I have discussed many times, the native category structure of Wikipedia (and all other commonly used KBs) leaves much to be desired. It is one of the major reasons we are re-organizing KB content using the UMBEL reference knowledge graph [1]. The ultimate requirement for the governing knowledge graph (ontology) is that it be logical, consistent and coherent. It is from this logical structure that we can provide the rich semsets [2] for semantic matches, make inferences, understand relatedness, and make disjointedness assertions. In the context of knowledge-based artificial intelligence (KBAI) applications [3], the disjointedness assertions are especially important to aiding the creation of negative training sets based on knowledge supervision.

Coherent and logical graphs first require natural groupings or classes of concepts and entity types by which to characterize the domain at hand, situated with respect to one another with testable relations. Entity types are further characterized by a similar graph of descriptive attributes. Concepts and entity types thus represent the nodes in the graph, with relations being the connecting infrastructure.

Going back at least to Aristotle, how to properly define and bound categories and concepts has been a topic of much philosophical discussion. If the nodes in our knowledge graph are not scoped and defined in a consistent way, then it is virtually impossible to construct a logical and coherent way to reason over this structure. This inconsistency is the root source of the problem that Wikipedia can not presently act as a computable knowledge graph, for example.

This article thus describes how Structured Dynamics informs its graph-construction efforts built around the notion of “natural classes.” Our use and notion of “natural classes” hews closely to how we understand the great American logician, Charles S. Peirce, came to define the concept [3]. Natural classes were a key underpinning to Peirce’s own efforts to provide a uniform classification system related to inquiry and the sciences.

Humanity’s Constant Effort to Define and Organize Our World

Aristotle set the foundational basis for understanding what we now call natural kinds and categories. The universal desire by all of us to be able to understand and describe our world has meant that philosophers have argued these splits and their bases ever since. In very broad terms we have realists, who believe things have independent order in the natural world and can be described as such; we have nominalists, who believe that humans provide the basis for how things are organized in part by how we name them; and we have idealists, or

[1]Email: mike@mkbergman.com

anti-realists, who believe “natural” classes are generalized ones that conform to human ideals of how the world is organized, but are not independently real [4]. These categories, too, shade into one another, such that these beliefs become strains in various degrees for how any one individual might be defined.The realist strain, also closely tied to the sciences and the scientific method, is what most guides the logic basis behind semantic technologies and SD’s view of how to organize the world.

Aristotle believed that the world could be characterized into categories, that categories were hierarchical in nature, and what defined a particular class or category was its essence, or the attributes that uniquely define what a given thing is. A mammal has the essences of being hairy, warm-blooded, and live births. These essences distinguish from other types of animals such as birds or reptiles or fishes or insects. Essential properties are different than accidental or artificial distinctions, such as whether a man has a beard or not or whether he is gray- or red-haired or of a certain age or country. A natural classification system is one that is based on these real differences and not artificial or single ones. Hierarchies arise from the shared generalizations of such essences amongst categories or classes. Under the Aristotelian approach, classification is the testing and logical clustering of such essences into more general or more specific categories of shared attributes. Because these essences are inherent to nature, natural clusterings are an expression of true relationships in the real world.

By the age of the Enlightenment, these long-held philosophies began to be questioned by some. Descartes famously grounded the perception of the world into innate ideas in the human mind. This philosophy built upon that of William of Ockham, who maintained the world is populated by individuals, and no such things as universals exist. In various guises, thinkers from Locke to Hume questioned a solely realistic organization of concepts in the world [5]. While there may be “natural kinds”, categorization is also an expression of the innate drive by humans to name and organize their world.

Relatedness of shared attributes can create ontological structures that enable inference and a host of graph analytics techniques for understanding meaning and connections. For such a structure to be coherent, the nodes (classes) of the structure should be as natural as possible, with uniformly applied relations defining the structure.

Thus, leaving behind metaphysical arguments, and relying solely on what is pragmatic, effectively built ontologies compel the use of a realistic viewpoint for how classes should be bounded and organized. Science and technology are producing knowledge at unprecedented amounts, and realism is the best approach for testing the trueness of new assertions. We think realism is the most efficacious approach to ontology design. One of the reasons that semantics are so important is that language used to capture the diversity of the real world must be able to be meaningfully related. Being explicit about the philosophy in how we construct ontologies helps decide sometimes sticky design questions.

Unnatural Classifications Instruct What is Natural

These points are not academic. The central failing, for example, of Wikipedia has been its category structure [7]. Categories have strayed from a natural classification scheme, and many categories are “artificial” in that they are compound or distinguished by a single attribute.“Compound” (or artificial) categories (such as `Films directed by Pedro Almodóvar` or `Ambassadors of the United States to Mexico`) are not “natural” categories, and including them in a logical evaluation only acts to confuse attributes from classification. To be sure, such existing categories should be decomposed into their attribute and concept components, but should not be included in constructing a schema of the domain.

“Artificial” categories may be identified in the Wikipedia category structure by both syntactical and heuristic signals. One syntactical rule is to look for the head of a title; one heuristic signal is to select out any category with prepositions. Across all rules, “compound” categories actually account for most of what is removed in order to produce “cleaned” categories.

We can combine these thoughts to show what a "cleaned" version of the Wikipedia category structure might look like. The 12/15/10 column in the table below reflects the approach used for determining the candidates for SuperTypes in the UMBEL ontology, last analyzed in 2010 [8]. The second column is from a current effort mapping Wikipedia to Cyc [9]:

	12/15/10	3/1/15
Total Categories	100%	100%
Administrative Categories	14%	15%
Orphaned Categories	10%	20%
Working Categories	76%	66%
"Artificial" Categories	44%	34%
Single Head	33%	23%
Plural Head		24%
"Clean" Categories	33%	46%

Two implications can be drawn from this table. First, without cleaning, there is considerable "noise" in the Wikipedia category structure, ***equivalent to about half to two-thirds of all categories***. Without cleaning these categories, any analysis or classification that ensues is fighting unnecessary noise and has likely introduced substantial assignment errors.

Second, the power that comes from a coherent schema of categories and concepts — especially inference and graph analysis — can not be applied to a structure that is not constructed along realistic lines. We can expand on this observation by bringing in our best logician on information, semeiosis and categories, Charles S. Peirce.

Peirce's Refined Arguments of a Natural Class

Peirce was the first, by my reading, who looked at the question of "natural classes" sufficient to provide design guidance, and which may be sometimes contraposed against what are called "artificial classes" (we tend to use the term "compound" classes instead). A "natural class" is a set with members that share the same set of attributes, though with different values (such as differences in age or hair color for humans, for example). Some of those attributes are also more essential to define the "type" of that class (such as humans being warm-blooded with live births and hair and use of symbolic languages). Artificial classes tend to only add one or a few shared attributes, and do not reflect the essence of the type [6].

The most comprehensive treatment of Peirce's thinking on natural classes was provided by Menno Hulswit in 1997. He first explains the genesis of Peirce's thinking [6]:

"The idea that things belong to natural kinds seems to involve a commitment to essentialism: what makes a thing a member of a particular natural kind is that it possesses a certain essential property (or a cluster of essential properties), a property both necessary and sufficient for a thing to belong to that kind."

"According to Mill, every thing in the world belongs to some natural class or real kind. Mill made a distinction between natural classes and non-natural or artificial classes (Mill did not use the latter term). The main difference is that the things that compose a natural class have innumerous properties in common, whereas the things that belong to an artificial class resemble one another in but a few respects."

"Accordingly, a natural or real class is defined as a class 'of which all the members owe their existence to a common final cause' (CP 1.204), or as a class the 'existence of whose members is due to a common and peculiar

final cause' (CP1.211). The final cause is described in this context as 'a common cause by virtue of which those things that have the essential characters of the class are enabled to exist' (CP 1.204)."

"Peirce concluded from these observations that the objects that belong to the same natural class, need not have all the characters that seem to belong to the class. After thus having criticized Mill, Peirce gave the following definition of natural class (or real kind):

> "*Any class which, in addition to its defining character, has another that is of permanent interest and is common and peculiar to its members, is destined to be conserved in that ultimate conception of the universe at which we aim, and is accordingly to be called 'real.' (CP6.384; 1901)"*

". . . natural classification of artificial objects is a classification according to the purpose for which they were made."

"The problem of natural kinds is important because it is inextricably linked to several philosophical notions, such as induction, universals, scientific realism, explanation, causation, and natural law."

This background sets up Hulswit's interpretation of then how Peirce's views on natural classification evolved [6]:

"Peirce's approach was broadly Aristotelian inasmuch as natural classification always concerns the form of things (which is that by virtue of which things are what they are) and not their matter. This entails that Peirce borrowed Aristotle's idea that the form was identical to the intrinsic final cause. Therefore it was obvious that natural classification concerns the final causes of the things. From the natural sciences, Peirce had learned that the forms of chemical substances and biological species are the expression of a particular internal structure. He recognized that it was precisely this internal structure that was the final cause by virtue of which the members of the natural class exist."

"Accordingly, Peirce's view may be summarized as follows: Things belong to the same natural class on account of a metaphysical essence and a number of class characters. The metaphysical essence is a general principle by virtue of which the members of the class have a tendency to behave in a specific way; this is what Peirce meant by final cause. This finality may be expressed in some sort of microstructure. The class characters which by themselves are neither necessary nor sufficient conditions for membership of a class, are nevertheless concomitant. In the case of a chair, the metaphysical essence is the purpose for which chairs are made, while its having chair-legs is a class character. The fuzziness of boundary lines between natural classes is due to the fuzziness of the class characters. Natural classes, though very real, are not existing entities; their reality is of the nature of possibility, not of actuality. The primary instances of natural classes are the objects of scientific taxonomy, such as elementary particles in physics, gold in chemistry, and species in biology, but also artificial objects and social classes."

"By denying that final causes are static, unchangeable entities, Peirce avoided the problems attached to classical essentialism. On the other hand, by eliminating arbitrariness, Peirce also avoided pluralistic anarchism. Though Peircean natural classes only come into being as a result of the abstractive and selective activities of the people who classify, they reflect objectively real general principles. Thus, there is not the slightest sense in which they are arbitrary: "there are artificial classifications in profusion, but [there is] only one natural classification" (C P 1.275; 1902)."

Importantly, note that "natural kinds" or "natural classes" are not limited to things only found in nature. Perice's semiotics (theory of signs) also recognizes "natural" distinctions in arenas such as social classes, the sciences, and man-made products [6]. Again, the key discriminators are the essences of things that distinguish them from other things, and the degree of sharing of attributes contains the basis for understanding relationships and hierarchies.

Natural Classes Can be Tested, Reasoned Over and Are Mutable

Though all of this sounds somewhat abstract and philosophical, these distinctions are not merely metaphysical. The ability to organize our representations of the world into natural classes also carries with it the ability to organize that world, reason over it, draw inferences from it, and truth test it. Indeed, as we may discover through knowledge acquisition or the scientific method, this world representation is itself mutable. Our understanding of species relationships, for example, has changed markedly, especially most recently, as the basis for our classifications shifts from morphology to DNA. Einstein's challenges to Newtonian physics similarly changed the "natural" way by which we need to organize our understanding of the world.

When we conjoin ideas such as Shannon's theory of information [10] with Peirce's sophisticated and nuanced theory of signs [11], other insights begin to emerge about how the natural classification of things ("information") can produce leveraged benefits. In linking these concepts together, de Tienne has provided some explanations for how Peirce's view of information relates to information theory and efficient information messaging and processing [12]:

"For a propositional term to be a predicate, it must have 'informed breadth', that is, it must be predicable of real things, 'with logical truth on the whole in a supposed state of information.' . . . . For a propositional term to be a subject, it must have 'informed depth', that is, it must have real characters that can be predicated of it also 'with logical truth on the whole in a supposed state of information'."

"Peirce indeed shows that induction, by enlarging the breadth of predicate terms, actually increases the depth of subject terms—by boldly generalizing the attribution of a character from selected objects to their collection—while hypothesis, by enlarging the depth of subject terms, actually increases the breadth of predicate terms—by boldly enlarging their attribution to new individuals. Both types of amplicative inferences thus generate information."

". . . information is not a mere sum of quantities, but a product, and this distinction harbors a profound insight. When Peirce began defining, in 1865, information as the multiplication of two logical quantities, breadth and depth (or connotation and denotation, or comprehension and extension), it was in recognition of the fact that information was itself a higher-order logical quantity not reducible to either multiplier or multiplicand. Unlike addition, multiplication changes dimensionality—at least when it is not reduced, as is often the case in schoolbooks, to a mere additive repetition. Information belongs to a different logical dimension, and this entails that, experientially, it manifests itself on a higher plane as well. Attributing a predicate to a subject within a judgment of experience is to acknowledge that the two multiplied ingredients, one the fruit of denotation, the other of connotation, in their very multiplication or copulative conjunction, engender a new kind of logical entity, one that is not merely a fruit or effect of their union, but one whose anticipation actually caused the union."

The essence of knowledge is that it is ever-growing and expandable. New insights bring new relations and new truths. The structures we use to represent this knowledge must themselves adapt and reflect the best of our current, testable understandings. Keeping in mind the need for all of our classes to be "natural" — that is, consistent with testable, knowable truth — is a key building block in how we should organize our knowledge graphs. Similar inspection can be applied to the relations used in the knowledge graph [13], but I will leave that discussion to another day.

Though hardly simple, the re-classification of Wikipedia's content into a structure based on "natural classes" will bring heretofore unseen capabilities in coherence and computability to the knowledge base. Similar benefits can be obtained from any knowledge base that is presently characterized by an unnatural structure.

We now have both tests and guidelines — granted, still being discerned from Peirce's writings or its logic — for what constitutes a "natural class". "Natural classes" are testable; we not only know it when we see it, we can systematize the use of them. In classifying a class as a "natural" one does entail aspects of judgment and world

view. But, so long as the logics and perspectives behind these decisions are consistent, I believe we can create computable knowledge graphs that cohere following these tests and guidelines.

Some may question whether any given structure is more "natural" than another one. But, through such guideposts as coherence, inference, testability and truthfulness, these structural arrangements are testable propositions. As Peirce, I think, would admonish us, failure to meet these tests are grounds for re-jiggering our structures and classes. In the end, coherence and computability become the hurdles that our knowledge graphs must clear in order to be reliable structures.

Acknowledgements

This article was originally posted on the *AI3::Adaptive Information* Web site at http://www.mkbergman.com/1876/natural-classes-in-the-knowledge-web/. This version has been edited and reformatted slightly for PDF distribution. We thank Cognonto Corporation for making this content freely available.

---

[1] For the latest release of UMBEL and its knowledge graph and associated links, see M.K. Bergman, 2015. "UMBEL version 1.20 Released," in *AI3:::Adaptive Information* blog, April 21, 2015.

[2] A **semset** is the use of a series of alternate labels and terms to describe a concept or entity. These alternatives include true synonyms, but may also be more expansive and include jargon, slang, acronyms or alternative terms that usage suggests refers to the same concept. See further http://wiki.opensemanticframework.org/index.php/Semset_Concept.

[3] I first discussed Charles S. Peirce at length in M.K. Bergman, 2012. "Give Me a Sign: What Do Things Mean on the Semantic Web," in *AI3:::Adaptive Information* blog, January 24, 2012.

[4] See, for example, John Michael Steiner, 2011. "An Anti-Realist Theory of Natural Kinds." PhD dissertation, University of Calgary, September 2011, 245 pp.

[5] Michael R. Ayers, 1981. "Locke versus Aristotle on Natural Kinds." *The Journal of Philosophy* (1981): 247-272.

[6] Menno Hulswit, 1997. "Peirce's Teleological Approach to Natural Classes," in *Transactions of the Charles S. Peirce Society* (1997): 722-772.

[7] See M.K. Bergman, 2015. "Shaping Wikipedia into a Computable Knowledge Base," in *AI3:::Adaptive Information* blog, March 31, 2015.

[8] Upper Mapping and Binding Exchange Layer (UMBEL) Specification, Annex G: UMBEL SuperTypes Documentation, *UMBEL.org*, retrieved February 16, 2015.

[9] Aleksander Smywinski-Pohl, Krzysztof Wróbel, Michael K. Bergman and Bartosz Ziółko, 2015. "cycloped.io: An Interoperable Framework for Web Knowledge Bases," manuscript in preparation.

[10] Claude E. Shannon, 1948. "A Mathematical Theory of Communication", *Bell System Technical Journal*, **27**: 379–423, 623-656, July, October, 1948. See http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[11] Charles Sanders Peirce, 1894. "What is in a Sign?", see http://www.iupui.edu/~peirce/ep/ep2/ep2book/ch02/ep2ch2.htm.

[12] André de Tienne, 2006. "Peirce's Logic of Information." *Seminario del Grupo de Estudios Peirceanos, Universidad de Navarra* 28 (2006).

[13] See, as one example, this discussion for the need for consistent and foundational relationship types, Giancarlo Guizzardi and Gerd Wagner, 2008. "What's in a Relationship: An Ontological Analysis" In *Conceptual Modeling-ER 2008*, pp. 83-97. Springer Berlin Heidelberg, 2008.