Cite as: Bergman, M. K. The Precepts. in *A Knowledge Representation Practionary: Guidelines Based on Charles Sanders Peirce* (ed. Bergman, M. K.) 85–104 (Springer International Publishing, 2018). doi:10.1007/978-3-319-98092-8_5

Official site: https://link.springer.com/book/10.1007/978-3-319-98092-8

Full-text: http://www.mkbergman.com/publications/akrp/chapter-5.pdf

Abstract: How we represent that reality has syntactic variation and ambiguities of a semantic nature can only be resolved by context. Once we resolve the source information, we need to organize it into ‘natural’ classes and relate those classes coherently and consistently to one another. Another precept is that knowledge graphs provide natural groupings of concepts and entity types to characterize the domain at hand, situated to one another with testable relations

5

THE PRECEPTS

To gain the opportunities in artificial intelligence and knowledge management, we need to look the world squarely in the eye and tackle realities as they exist. As a philosopher, Charles S. Peirce was a confirmed and staunch defender of realism, though he was also an idealist in his belief that truth, while not perhaps knowable in its absolute limits, could be increasingly discovered through the application of logic and the scientific method. Pragmatism is the way forward to approach this ideal.

The world is a messy place. Not only is it complicated and richly diverse, but our ways of describing and understanding it are made more complex by differences in language and culture. We know the world is interconnected and interdependent. Effects of one change can propagate into subtle and unforeseen consequences. Not only is the world always changing, but so is our understanding of what exists in the world and how it affects and is affected by everything else. This continuous flux means we are always uncertain to a degree about how the world works and the dynamics of its working. Through education and research we continually strive to learn more about the world, but often in that process find what we thought was true is no longer so and even our human existence is modifying our world in manifest ways.[1]

Knowledge is very similar to this nature of the world. We find that knowledge is never complete and we can see it anywhere and everywhere. We capture and codify knowledge in structured, semi-structured and unstructured forms, ranging from 'soft' to 'hard' information. We find that the structure of knowledge evolves with the incorporation of more information. We often see that knowledge is not absolute, but contextual. That does not mean truth does not exist; rather knowledge should be coherent, to reflect a logical consistency and structure that comports with our observations about the physical world. Knowledge, like the world, is continually changing; we thus must adapt to what we observe and learn.

Chapter 3 pointed to the importance of information to economic growth. We saw the breakpoint accelerations in growth tied to historical changes in the cost and access to information. Future generations will surely come to see the Internet phenomenon as one of those transitions. Massive storehouses of information, under free and open licenses, are available at our fingertips. None of these sources were designed for interoperability at a concept or knowledge level, and each has its context, format, and terminology. Here is a practically unlimited source of useful information that, by applying our approaches and principles to semantic technologies and interoperability, we can tap for digital reasoning and learning.

To tap this storehouse of information, to connect and make the information usable with other information, we need to understand what makes that information in its raw form a Tower of Babel. To overcome these differences we need to embrace some premises — or *precepts* — about how our information exists in its native forms, and then to adopt still further propositions for how to put that information on a common footing. These precepts relate to the nature of data, semantic heterogeneities in what that data means, and how we organize and classify that data. These precepts help set the ground rules for our actions going forward.

EQUAL CLASS DATA CITIZENS

Knowledge representation, by our definition, operates in an electronic medium with messages conveyed in bits, which makes all information represented in the system as data. To deal in the realm of knowledge and belief, the purpose of our KR systems, we must be able to ingest and process any electronic data in any form that can contribute to our knowledge.* We include any digital information artifact in this category, including ‘soft’ or ‘hard’ information, social information, information of varying certainty, and information of diverse provenance. We thus define *content* as information that has the potential to contribute to knowledge.

These variations are what would be called syntactic, or the structure or form of the information, though content ambiguities also lead to an entirely different plane of differences, those of a semantic nature. Whatever these differences of structure, format or content, as long as the information represented is a possible contributor to knowledge, we must be able to ingest and process it. Knowledge management systems must treat all data forms with a potential to contribute to knowledge as equal class citizens.

* While streaming media alone does not meet this definition, transcripts or tags associated with the content do.

The Structural View

A favorite, and I think useful, split of content is according to its native structure; that is, the structure it assumes when created for its primary purpose. One of these groupings is *structured data*, what we most often think of when we hear the term 'data.' This classification is where the information presented is according to a defined data model, commonly found in relational databases or other forms of tabular data, such as even an electronic spreadsheet. This information includes any managed by database management software, but it can also be as simple as an HTML table for the Web. We can model, organize, form, and format structured data in ways that are easy for us to manipulate. Much of our current know-how related to data and its management comes from our decades-long experience with structured data.

The second grouping of content is *unstructured data*, mostly consisting of text, which lacks an explicit data model or schema. (But it does comport with the 'structure' of natural languages.) All documents and output from word processors or editors fall into this category, as do transcriptions of talking or speech. For decades, researchers have estimated the amount of information within an enterprise embedded in text documents to approximate 80% of available information; some recent estimates put that contribution at 90%.[2] Whatever the number, the percentage of information in documents represents the preponderance of what might be useful for knowledge purposes within the organization.

The third grouping of content is thus *semi-structured data*, which is of more recent vintage. This category of content does not conform to a formal tabular or structural data model but gets its 'semi-structured' nature by embedding tags or other markers to denote fields within the content. We obtain it from unstructured data via data mining or information extraction. Separate annotations not embedded within the text, as is the case for *metadata*, are also part of this grouping. Markup languages embedded in text are a common form of such sources.

Semi-structured data provides something of a 'middle ground' between structured and unstructured sources. Semi-structured data models are sometimes called 'self-describing' (or schema-less).[3] The first known definition of semi-structured data dates to 1993 by Peter Schäuble.[4] More current usage also includes the notion of labeled graphs or trees with the data stored at the leaves, with the schema information contained in the edge labels of the graph. Semi-structured representations also lend themselves well to data exchange or the integration of heterogeneous data sources. Another nice aspect of semi-structured formats is that they are readable as text, with a structure that can be understood and assigned by non-programmers without dedicated IT staff. Semi-structured data is the preferred form for annotations.

To date, we have good processing engines for specific semi-structured forms, such as rendering HTML in a Web browser or reading XML data sources, but inadequate engines for combining different forms of semi-structured data.[5] Moreover, semi-structured data is the basis for including unstructured text with structured data, but we still have the issues of extracting structure from various formats.

The Formats View

Broad categorizations by content, while useful for generalizing, mask the ways we can express these unstructured, semi-structured, and structured forms in various file formats. Since no 'official' repository for file formats exists, it is impossible to know how many different flavors of formats exist in the wild. The most extensive reference that I have found, kept on Wikipedia, lists nearly 1500 different file formats from AAC (audio coding) to zoo (file compression).[9] Perhaps this sounds worse than it should because these file formats span entire application areas from documents to archives to audio or video or gaming. Further, skewed power-law distributions mean only a fraction of these formats account for most uses. Individual application areas, such as word processing or spreadsheets, have merely scores or hundreds of formats. What we experience in the wild is also the subset of more popular formats, though in a medium as broad as the Web, Google has found tens of thousands of individual schemata.[7] Single enterprises may need only deal with a few score common formats, rather than thousands, with perhaps only a few dominant formats in given application areas. Word processors, for example, might be mandated or standardized. Still, even in this area, much readable text in multiple other formats is available.

Structured or semi-structured data formats also have a schema, in addition to the serialization formats used for transmittal. Some markup languages, such as HTML or Markdown, have embedded tags that instruct how to render Web pages or guide the user interface. Other markup languages, such as fielded text, structured text, simple declarative language (SDL), or more recently YAML or its simpler cousin JSON, have become more widely adopted and supported by formal specifications, tools or APIs. Many prefer JSON, for example, as a form for Web applications. Some formats, like microformats or BibTeX records, rely less on syntax conventions and may use reserved keywords (such as AUTHOR or TITLE) to signal the key for the *key-value pair*.

These various forms, sometimes well specified with APIs and sometimes almost *ad hoc* as in spreadsheet listings, are what we call 'structs.' *Structs* can all be displayed as text and have, at a minimum, explicit or inferrable key-value pairs to convey data relationships and attributes, with data types and values often noted by various white space, delimiter (such as angle brackets) or other text conventions. Some of these simple formats have been more successful than others, though none have achieved market dominance. Few universal principles have emerged as to syntax or format. One positive is that most of these various *struct* forms are easy for casual users to understand and easy for domain experts to write.

The sheer number of file formats one may encounter in the wild (including within the single organization) is such that pairwise translators between forms are not combinatorially possible. The only way to handle the diversity of forms and formats is to establish one or a limited few *canonical* formats and to translate wild forms to those formats. This scalable approach to federation is a central topic of *Chapter 9*.

The Content View

Refer to Jimmy Johnson by his name, and you might be referring to a former football coach, a NASCAR driver, a former boxing champ, a blues guitarist, or perhaps even a plumber in your hometown. Alternatively, perhaps your Jimmy is none of these individuals. The label 'Jimmy Johnson' is insufficient to establish identity. As another example, let's take the seemingly simple idea of 'cats.' In one source, the focus might be on house cats, in a second domestic cats, and in a third, cats as pets. Are these ideas the same thing? Now, let's bring in some taxonomic information about the cat family, the Felidae. We have now expanded the idea of 'cats' to include lynxes, tigers, lions, cougars and many other kinds of cats, domestic and wild (and, also extinct). The 'cat' label used alone clearly fails us miserably here.

As a third example, let's take the concept or idea of the named entity of Great Britain:

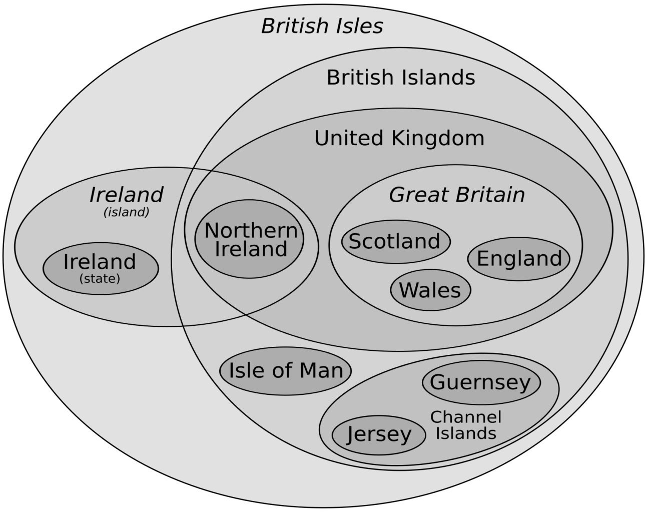

Figure 5-1: Names Can Be Complicated[16]

Depending on usage and context, Great Britain can refer to quite different scopes and things. In one sense, Great Britain is an island. In a political sense, Great Britain can comprise the territory of England, Scotland, and Wales. Even more, precise understandings of that political grouping may include some outlying islands such as the Isle of Wight, Anglesey, the Isles of Scilly, the Hebrides, and the island groups of Orkney and Shetland. Sometimes the Isle of Man and the Channel Islands, which are not part of the United Kingdom, are included in error in that political grouping. Then, in another context, Great Britain may also include Northern Ireland, since the two countries sometimes combine their sports teams. These, plus other confusions, can mean quite different things when referring to 'Great Britain' as the Venn diagram of possibilities shows us in *Figure 5-1*.[*]

Even with the same label, items in different information sources can refer to the same thing, but may not be the same thing or may define it with a different scope and content. *Ambiguity* is one source for such error, as our examples show. If we incorrectly identify the object, then connections can get drawn that are in error, which is why *disambiguation* is such a big deal in knowledge systems. In broad terms, these mismatches can be due to structure, domain, data or language, with many nuances within each type.

Data without context and relationships are meaningless. Logic and consistency almost by definition imply the application of a uniform perspective, a single worldview. Multiple authors making contributions without a common frame of reference or viewpoint are unable to bring this consistency of perspective. The `sameAs` approach used to connect items in many current Web systems when they ignore such heterogeneities, makes as little sense as talking about the plumber using facts drawn from the blues guitarist. Even if we can overcome the syntactic and format differences already discussed, we still face the hurdle of bridging the semantics of the data federation pyramid shown in *Figure 5-1*.

ADDRESSING SEMANTIC HETEROGENEITY

The *idea of* something — that is, its *meaning* — is conveyed by how we define that something, the context in which we use the various tokens (terms) for that something, and in the variety of words or labels we apply to that thing. The label alone is not enough. We convey the *idea of a parrot* by our understanding of what the name *parrot* means. In languages other than English, the same idea of parrot may be conveyed by the terms Papagei, perroquet, loro, попугай, or オウム, depending on the native language. The idea of the ‘*United States*,‘ even just in English, may be conveyed with labels ranging from *America* to *US*, *USA*, *U.S.A.*, *Amerika*, *Uncle Sam*, or even the *Great Satan*. What these examples illustrate is that a single term is more often not the only way to refer to something, and a given token may mean vastly different things depending on context. The oft-heard phrase, ‘things, not strings,’ captures this underlying fact.[8]

Sources of Semantic Heterogeneity

Our understanding of the patterns in semantic heterogeneities — for which we need to account in the design of our knowledge systems explicitly — is pretty mature. We see confusion potentially arising from multiple terms for a single thing;[†] single terms applying to numerous things; terms whose meaning derives from context; how we characterize things; how we relate things; how we indicate surety or confidence; how we point to things; and, how to annotate things.

[*] These associations also vary over time, again well evidenced by the scope of ‘Great Britain.’

[†] Though true synonyms are rare, our practical interest is to capture alternate labels for the same thing.

Pluempitiwiriyawej and Hammer provided one of the first comprehensive schemes for classifying semantic heterogeneities.[9] I have used and added to this schema over many years. By decomposing this space into its various sources of semantic heterogeneities — as well as the work required to provide for such functions as search, disambiguation, mapping, and transformations — we can begin to understand how all of these components can work together to help achieve data interoperability.

The following *Table 5-1* shows more than 40 sources of semantic heterogeneities, structurally organized, each of which is a possible impediment to get data to interoperate across sources:

Class	*Category*	*Subcategory*	*Examples*	*Type* [12]
LANGUAGE	Encoding	Ingest Encoding Mismatch	For example, ANSI *v* UTF-8[13]	Concept
		Ingest Encoding Lacking	Mis-recognition of tokens because not being parsed with the proper encoding [13]	Concept
		Query Encoding Mismatch	For example, ANSI *v* UTF-8 in search [13]	Concept
		Query Encoding Lacking	Mis-recognition of search tokens because not being parsed with the proper encoding [13]	Concept
	Languages	Script Mismatch	Variations in how parsers handle, say, stemming, white spaces or hyphens	Concept
		Parsing / Morphological Analysis Errors (many)	Arabic languages (right-to-left) *v* Romance languages (left-to-right)	Concept
		Syntactical Errors (many)	Ambiguous sentence references, such as *I'm glad I'm a man, and so is Lola* (*Lola* by **Ray Davies** and the **Kinks**)	Concept
		Semantics Errors (many)	River *bank v* money *bank v* billiards *bank* shot	Concept
CONCEPTUAL	Naming	Case Sensitivity	Uppercase *v* lower case *v* Camel case	Concept
		Synonyms	United States *v* USA *v* America *v* Uncle Sam v Great Satan	Concept
		Acronyms	United States *v* USA *v* US	Concept
		Homonyms	Such as when the same name refers to more than one concept, such as Name referring to a person *v* Name referring to a book	Concept
		Misspellings	As stated	Concept

Class	*Category*	*Subcategory*	*Examples*	*Type* [12]
	Generalization / Specialization		When single items in one schema are related to multiple items in another schema or vice versa. For example, one schema may refer to 'phone,' but the other schema has multiple elements such as 'home phone,' 'work phone' and 'cell phone'	Concept
	Aggregation	Intra-aggregation	When the same population is divided differently (such as Census *v* Federal regions for states, England *v* Great Britain *v* United Kingdom, or full person names *v* first-middle-last)	Concept
		Inter-aggregation	May occur when we include sums or counts as set members	Concept
CONCEPTUAL	Internal Path Discrepancy		Can arise from different source-target retrieval paths in two different schemas (for example, hierarchical structures where the elements are different levels of remove)	Concept
	Missing Item	Content Discrepancy	Differences in set enumerations or including items or not (say, US territories) in a listing of US states	Concept
		Missing Content	Differences in scope coverage between two or more datasets for the same concept	Concept
		Attribute List Discrepancy	Differences in attribute completeness between two or more datasets	Attribute
		Missing Attribute	Differences in scope coverage between two or more datasets for the same attribute	Attribute
	Item Equivalence		When we assert two types (classes or sets) as being the same when the scope and reference are not (for example, **Berlin** the city *v* **Berlin** the official city-state)	Concept
			When we assert two individuals as being the same when they are distinct (for example, **John Kennedy** the president *v* **John Kennedy** the aircraft carrier)	Attribute
	Type Mismatch		When we characterize the	Attribute

Class	*Category*	*Subcategory*	*Examples*	*Type* [12]
			same item by different types, such as a person typed as an animal *v* human being *v* person	
	Constraint Mismatch		When attributes referring to the same thing have different cardinalities or disjointedness assertions	Attribute
DOMAIN	Schematic Discrepancy	Element-value to Element-label Mapping	One of four errors that may occur when attribute names or values may not be completely unambiguous.	Attribute
		Attribute-value to Element-label Mapping		Attribute
		Element-value to Attribute-label Mapping		Attribute
		Attribute-value to Attribute-label Mapping		Attribute
	Scale or Units	Measurement Type	Differences, say, in the metric *v* English measurement systems, or currencies	Attribute
		Units	Differences, say, in meters *v* centimeters *v* millimeters	Attribute
DOMAIN	Precision		For example, a value of 4.1 inches in one dataset *v* 4.106 in another dataset	Attribute
	Data Representation	Primitive Data Type	Confusion often arises in the use of literals *v* URIs *v* object types	Attribute
		Data Format	Delimiting decimals by period *v* commas; various date formats; using exponents or aggregate units (such as thousands or millions)	Attribute
DATA	Naming	Case Sensitivity	Uppercase *v* lower case *v* Camel case	Attribute
		Synonyms	For example, centimeters *v* cm	Attribute
		Acronyms	For example, currency symbols *v* currency names	Attribute
		Homonyms	Such as when the same name refers to more than one attribute, such as Name referring to a person *v* Name referring to a book	Attribute
		Misspellings	As stated	Attribute
	ID Mismatch or Missing ID		URIs can be a particular problem here, due to actual mismatches but also use of names-	Attribute

Class	*Category*	*Subcategory*	*Examples*	*Type* [12]
			paces or not and truncated URIs	
	Missing Data		A common problem, more concerning with closed world approaches than with open world ones	Attribute
	Element Ordering		Set members can be ordered or unordered, and if ordered, the sequences of individual members or values can differ	Attribute

Table 5-1: Sources of Semantic Heterogeneities

We have assigned these structural aspects to one of two types: a) those that may arise from the *conceptual* differences between sources (mostly from schema differences); and b) those due to value and *attribute* discrepancies (data). The table also provides examples of what each of these categories of heterogeneities means.

This listing is a reasonably comprehensive view of what is involved in getting things to talk together (*semantic agreement*). Fortunately, via the adoption of standard syntactic protocols and semantic languages, means for managing many of these possible heterogeneities are handled in the background when complying with their rules (axioms) or language constructs. That still leaves us with the heterogeneities associated with human communications and how to measure the attributes of things.

From the conceptual to actual data, then, we see differences in perspective, vocabularies, measures, and conventions. Some of these differences and heterogeneities are intrinsic to the nature of the data at hand. Some of these heterogeneities also arise from the basis and connections asserted between datasets. Only by systematically understanding these sources of heterogeneity — and then explicitly addressing them — can we begin to try to put disparate information on a common footing. Only by reconciling differences can we start to get data to interoperate.

Role of Semantic Technologies

The first advantage of semantic technologies is that all kinds of information are unified. No matter what information you consider, any content type may become a 'first-class citizen.' For really the first time, we can put all kinds of information ranging from traditional databases and spreadsheets (*structured*) to markup, Web pages, XML and data messages (*semi-structured*), and then on to documents and text (*unstructured*) or multimedia (via *metadata*) on a level playing field. These data, now all treated on an equal footing, can be searched and retrieved by a variety of techniques. These range from SQL, standard text search, or SPARQL, depending on content type. This unique combination enables us to fulfill all of the aspects of findability — find, discover, navigate. Because of the diversity of search options available, we can vary and optimize search results depending on circumstance and needs. Because all content is represented either as a type of thing, an individual thing, or the relationships between those things, we may use these classifiers for faceting or grouping. Further, the connections put all things in context, useful to ensure results are relevant and disambiguated.

What works efficiently for transactions and accounting is a poor choice for knowledge problems. Traditional relational databases work best with structured data; are inflexible and fragile when the nature (schema) of the world changes; and thus require constant (and expensive) re-architecting in the face of new knowledge or new relationships. Conversely, for semantic technologies, we describe things and their relationships based on the 'idea of the thing,' not limited to keywords. Thus, we can describe and find things using alternative terms, synonyms, acronyms or jargon. We can add on or extend semantic vocabularies without altering what we have already asserted, assuming the prior assertions still hold true.

We should use semantic technologies instead of conventional information technologies in the areas of knowledge representation (KR) and knowledge management (KM). Semantic technologies are orthogonal to some other current technologies, including cloud computing and big data. Semantic technologies are not limited to open data: they are equivalently useful to private or proprietary data. Semantic technologies do not imply some grand, shared schema for organizing all information, though, at some levels, that is extremely useful. Semantic technologies are not 'one ring to rule them all,' but rather a way to capture the worldviews of particular domains and groups of stakeholders. Semantic technologies appropriately done are not a replacement for existing information technologies, but rather an added layer that can leverage those assets for interoperability and to overcome the semantic barriers between existing information silos. These very same semantic technologies also provide the proper representational basis for symbol-based machine learning and intelligence.

Semantic technologies give us the basis for understanding differences in meaning across sources, specifically geared to address differences in real-world usage and context. These semantic tools are essential for providing common bases for relating structured data across various sources and contexts. These same semantic tools are also the basis by which we can determine what unstructured content 'means,' thus providing the structured data tags that also enable us to relate documents to conventional data sources using semi-structured data. Semantic technologies are therefore the enablers for making information understandable to both humans and machines across sources.

Semantic technologies expressly address these heterogeneities, some more strongly in some areas than others. However, to capture the scope of the heterogeneities listed, we need the technologies to mimic aspects of human language, symbology, and logic. We express ourselves via the equation and the document, not to mention jumping up and down and gesticulating. By accounting explicitly for the relationships between things, we can use semantic technologies to better capture context, essential for navigation and the reduction of ambiguity. We can use the richness of relationships to group, classify, filter, or find things. The basic assertion in our semantic languages declares relationships between and for things. These statements, when combined with the objects of some statements being the subjects of others, leads to a graph structure (see *Chapter 1*). We may apply various logics based on the nature of our declarations to compute over the structure and understand or infer relationships between things. We can use the graph structures for novel traversal mechanisms and network analysis. No other information structure provides these unique advantages.

Semantics and Graph Structures

The graph structures of semantic schema mean that any node can become an entry point to the knowledge space for discovery. The traversal of information relationships occurs from the selection of predicates or properties that we use to create this graph structure in the first place. This richness of characterization and objects also means we can query or traverse this space in multiple languages or via the full spectrum by which we describe or characterize things. Semantic-based knowledge graphs are potentially an explosion of richness in characterization and how those characterizations get made and referred to by any stakeholder.[10] We enable the user community to determine our search structures, rather than some group of designers or information architects. It should not be surprising that search offers one of the quickest and most visible paths to gain the benefits of semantic technologies.

Existing IT assets represent massive sunk costs, legacy knowledge and expertise, and (often) stakeholder consensus. These systems are still mostly stovepiped. Strategies that counsel replacing existing IT systems risk wasting existing assets. We are better served to leverage the value already embodied in these systems while promoting interoperability and integration. The beauty of semantic technologies — adequately designed and deployed in a Web-oriented architecture — is that a thin interoperability layer may be placed over existing IT assets to achieve these aims. We can use the knowledge graph structure to provide the semantic mappings between schema, while we use a Web service framework to convert sources to the canonical data model. Via these approaches, we may preserve prior investments in knowledge, information, and IT assets while enabling interoperability. The existing systems can continue to provide the functionality as initially deployed. Meanwhile, we may expose and integrate the KR-related aspects with other knowledge assets on the physical network. Being able to manage semantic heterogeneity is the kickstarter to this process.

CARVING NATURE AT THE JOINTS

The embracing of semantics and the languages to express them is but the prerequisite. Once we decide the rules of the game, we need to populate our domain. That means we need to capture the concepts, instances, attributes, and relations of our domain. This capturing forms our vocabulary, and how we group, classify and type that vocabulary should reflect the reality of our domain and how we organize it.

Stated in the abstract this sounds like a tall order. However, we help fulfill this order if we seek to organize our domain in the most realistic way possible, what Plato, speaking as Socrates in the dialog with Phaedrus, says:[11]

> SOCRATES
>
> *It seems to me that the discourse was, as a whole, really sportive jest; but in these chance utterances were involved two principles, the essence of which it would be gratifying to learn, if art could teach it.*
>
> PHAEDRUS
>
> *What principles?*
>
> SOCRATES
>
> *That of perceiving and bringing together in one idea the scattered particulars, that one may make clear by definition the particular thing which he wishes to explain; just as now, in speaking of Love, we said what he is and defined it, whether well or ill. Certainly by this means the discourse acquired clearness and consistency.*
>
> PHAEDRUS
>
> *And what is the other principle, Socrates?*
>
> SOCRATES
>
> *That of dividing things again by classes, where the natural joints are, and not trying to break any part, after the manner of a bad carver. As our two discourses just now assumed one common principle, unreason, and then, just as the body, which is one, is naturally divisible into two, right and left, with parts called by the same names, so our two discourses conceived of madness as naturally one principle within us, and one discourse, cutting off the left-hand part, continued to divide this until it found among its parts a sort of left-handed love, which it very justly reviled, but the other discourse, leading us to the right-hand part of madness, found a love having the same name as the first, but divine, which it held up to view and praised as the author of our greatest blessings.*
>
> PHAEDRUS
>
> *Very true.*
>
> SOCRATES
>
> *Now I myself, Phaedrus, am a lover of these processes of division and bringing together, as aids to speech and thought; and if I think any other man is able to see things that can naturally be collected into one and divided into many, him I follow after and «walk in his footsteps as if he were a god.»*

The idea of 'carving nature at the joints' is a mindset we can apply to all of the major divisions in our vocabulary; namely, things, concepts, relations, and attributes.

Forming 'Natural' Classes

As we see, going back at least to Plato and Aristotle, how to properly define and bound categories and concepts have been a topic of much philosophical discussion. If we do not scope the organization of our knowledge and define it consistently, then it is virtually impossible to construct a logical and coherent way to reason over this structure. Aristotle set the foundational basis for understanding what we now call natural kinds and categories (or 'classes'). The universal desire to understand and describe our world has meant that philosophers have argued these splits and their bases ever since. We can place these philosophical arguments into three broad camps. First, we have *realists*, who believe things have independent order and existence in the natural world, apart from thought. Second, we have *nominalists*, who believe that humans provide the basis for how things are organized in part by how we name them. Third, we have *idealists*, or 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.[12] These categories shade into one another, such that these beliefs become strains in various degrees for how any one philosophy might be defined.

The realist strain, also closely tied to the sciences and the scientific method, is what most guides the logical basis of semantic technologies and our view of how to organize the world. Science and technology are producing knowledge in unprecedented amounts, and realism is the best approach for testing the trueness of new assertions. We think realism is the most efficacious approach to knowledge representation designs. Being explicit about the philosophy in how we construct our knowledge representations helps decide sometimes sticky design questions, as we will see multiple times throughout this book.

Aristotle believed that the world fits 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 mammals from other types of animals such as birds or reptiles or fishes or insects. Essential properties are different from 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. We base a natural classification system on real differences of character 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 real relationships in the real world, often hierarchical in structure.

By the age of the Enlightenment, some began to question these long-held philosophies. Descartes famously grounded the perception of the world into innate ideas in the human mind. Descartes' philosophy, built upon that of William of Ockham of Occam's razor fame, maintained individuals populate the world; no such things as universals exist. In various guises, thinkers from Locke to Hume questioned a solely realistic organization of concepts in the world.[13] While there may be 'natural kinds,' categorization is also an expression of the innate drive by humans to name and organize their world, was the dominant view of these emerging nominalists.

Charles S. Peirce started a mighty swing back to realism. He was the first, by my reading, who looked at the question of 'natural classes' sufficient to provide design guidance, and which may sometimes be contraposed against what some call 'artificial classes' (we also tend to use the term 'compound' classes). Natural classes were a key underpinning to Peirce's own efforts to provide a uniform classification system related to inquiry and the sciences. A *natural class* is a set of 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 only to add one or a few shared attributes and do not reflect the essence of the type.[18] Our use and notion of 'natural classes' hews closely to how Peirce understood the concept:

> "So then, a natural class being a family whose members are the sole offspring and vehicles of one idea, from which they derive their peculiar faculty, to classify by abstract definitions is simply a sure means of avoiding a natural classification. I am not decrying definitions. I have a lively sense of their great value in science. I only say that it should not be by means of definitions that one should seek to find natural classes. When the classes have been found, then it is proper to try to define them; and one may even, with great caution and reserve, allow the definitions to lead us to turn back and see whether our classes ought not to have their boundaries differently drawn. After all, boundary lines in some cases can only be artificial, although the classes are natural ...." (EP 2:125)

Peirce's ideas of a natural kind appear closely tied to his realism:

> "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.'" (1901, CP 6.384)

Another guideline that Peirce provides is that *intension* is also a means for determining a natural classification:

> "The descriptive definition of a natural class, according to what I have been saying, is not the essence of it. It is only an enumeration of tests by which the class may be recognized in any one of its members. A description of a natural class must be founded upon samples of it or typical examples." (1902, CP 1.223)

Peirce greatly admired the natural classification systems of Louis Agassiz and used animal lineages in many of his examples. He was a strong proponent of natural classification. Though we have replaced the morphological basis for classifying organisms in Peirce's day with genetic ones, Peirce would surely support this new knowledge, since he grounded his philosophy on a triad of primitive unary, binary and tertiary relations, bound together in a logical sign process seeking truth.

For example, natural class instances, which are by definition intensional due to the *differentia* that comprises their class, may be declared by assignment to a class type. Once we define a type such as a hairless mammal that walks in a bipedal manner as a *human*, we can after that assign individual people to that class type and thereby infer human properties (or characteristics). We need not specify all possible human properties per individual under a strictly intensional approach nor enumerate all human individuals under a strictly extensional approach. We can let the use of type assignments bridge this divide. We can also see that, depending on context, we may want to speak about *human* as a class (type) subsuming individual people or to speak about *human* as an instance with particular kinds of properties (attributes). I discuss further this ‘punning’ metamodeling technique in *Chapter 9*.

Peirce’s concept of ‘natural kinds’ or ‘natural classes’ is not limited to things only found in nature. Peirce's semiotics (theory of signs) also recognizes ‘natural’ distinctions in areas such as social classes, the sciences, and human-made products.[14] These distinctions are important because they affirm essences and realities in the external world. A ‘natural’ classification is not limited to the animate. ‘Natural’ classification is premised on reason and subject to testing. 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.

Menno Hulswit is one of the scholars who has studied Peirce’s concept of ‘natural classes’ most closely.[18] As he has observed:

> “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.” (p. 759)
>
> “… 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.' (1902, CP 1.275)" (pp. 765-6)

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 physical world.

A Mindset for Categorization

These points are not academic. The central weakness, for example, that I have noted for Wikipedia over many years has been its category structure. Category inconsistencies are the root source of the problem that Wikipedia can not presently act as a computable knowledge graph.* Categories often do not conform to 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, we should decompose such existing categories into their attribute and concept components, and possibly only include the decomposed versions (if then) when 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 account for most of what we remove to produce 'cleaned' categories. Including administrative and other problem categories, perhaps half to two-thirds of Wikipedia's categories do ***not*** meet the definition of natural categories, though Wikipedia's editors continue to make improvements.[15] Independent actors have staged and processed Wikipedia multiple times to overcome these limits to create usable knowledge bases.

Whatever the target for the categorization effort may be, Peirce put forward some general execution steps:

> "... introduce the monadic idea of »first« at the very outset. To get at the idea of a monad, and especially to make it an accurate and clear conception, it is necessary to

* Some reviewers have suggested the issue is a matter of scale. While I agree large scale causes its own challenges, I believe the problem is one more of coherence and lack of consistency.

> begin with the idea of a triad and find the monad-idea involved in it. But this is only a scaffolding necessary during the process of constructing the conception. When the conception has been constructed, the scaffolding may be removed, and the monad-idea will be there in all its abstract perfection. According to the path here pursued from monad to triad, from monadic triads to triadic triads, etc., we do not progress by logical involution — we do not say the monad involves a dyad — but we pursue a path of evolution. That is to say, we say that to carry out and perfect the monad, we need next a dyad. This seems to be a vague method when stated in general terms; but in each case, it turns out that deep study of each conception in all its features brings a clear perception that precisely a given next conception is called for." (1896, CP 1.490)

This quote is at the root of Peirce's views concerning the universal categories, the main topic of the next chapter. Triads figure prominently in this thinking. As we weave the various threads in Peirce's philosophy together, we also come to see the logic of how the three components of inquiry work in a similar manner to categorization, itself just a more structured view of what Peirce discussed as a generalization. What we learn from Peirce in this investigation is that categorization, thankfully, is a knowledge representation task, that we can approach logically and systematically. We can adopt a categorical mindset about how to think of the world. The assignments should be defensible, but we should also be ready to change them when faced with better evidence or logic. We learn more about how to think through categorization in *Chapter 6*.

Connections Create Graphs

When representing knowledge, more things and concepts get drawn into consideration. In turn, the relationships of these things lead to connections between them to capture the inherent interdependence and linkages of the world. As still more things get considered, we make and proliferate more connections. This process naturally leads to a graph structure, with the things in the graphs represented as nodes and the relationships between them represented as connecting edges.* More things and more connections lead to more structure. Insofar as this structure and its connections are coherent, the natural structure of the knowledge graph itself can help lead to more knowledge and understanding.

Coherent and logical graphs first require natural groupings or classes of concepts and entity types by which to characterize the domain at hand, situated to one another with testable relations. We characterize entity types with a similar graph of descriptive attributes. Concepts and entity types thus represent the nodes in the graph, with relations being the connecting infrastructure. Relatedness of shared attributes or types of relations can also 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 also be as natural as possible, applying the same categorization approaches.

Unlike traditional data tables, graphs have some inherent benefits, particularly

* See *Figure 1-3*.

for knowledge representations. They provide:

- A coherent way to navigate the knowledge space;
- Flexible entry points for each user to access that knowledge (since every node or relation is a potential starting point);
- Inferencing and reasoning structures about the space;
- Connections to related information;
- Ability to connect to any form of information;
- Concept mapping, and thus the ability to integrate external content;
- A framework to disambiguate concepts based on relations and context; and
- A common vocabulary to drive content 'tagging.'

Graphs are the natural structures for knowledge domains if they follow a 'natural' classification and we test them for coherence. Once built, graphs offer some analytical capabilities not available through traditional means of information structure. Graph analysis is a rapidly emerging field, but we are already able to gauge some unique measures of knowledge domains, such as influence, relatedness, proximity, centrality, inference, clustering, shortest paths, and diffusion. As science is coming to appreciate, graphs can represent any extant structure or schema. The universal character of graphs makes them an attractive target for many analytic tools.

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 'natural' classes — that is, consistent with testable, knowable truth — is a building block in how we should organize our knowledge graphs. Through such guideposts as coherence, inference, and truthfulness, these structural arrangements become testable propositions. As Peirce, I think, would admonish us, failure to meet these tests is grounds for re-jiggering our structures and classes. In the end, coherence and computability become the hurdles that our knowledge graphs must clear to become reliable structures.

Chapter Notes

1. Some material in this chapter was drawn from the author's prior articles at the *AI3:::Adaptive Information* blog: "What is Linked Data?" (Jun 2008); "When is Content Coherent?" (Jul 2008); "'Structs': Naïve Data Formats and the ABox" (Jan 2009); "The Law of Linked Data" (Oct 2009); "When Linked Data Rules Fail" (Nov 2009); "The Bipolar Disorder of Linked Data" (Apr 2010); "Practical P-P-P-Problems with Linked Data" (Oct 2010); "What is Structure?" (May 2012); "The Rationale for Semantic Technologies" (Jul 2012); "Making Text a First-Class Citizen" (Jan 2013); "Three Leading Arguments for Semantic Technologies" (Jan 2013); "Seven Arguments for Semantic Technologies" (Feb 2013); "The Primacy of Search in the Semantic Enterprise" (Feb 2013); "'Natural Classes' in the Knowledge Web" (Jul 2015); "SWEETpedia" (available at http://www.mkbergman.com/sweetpedia/).
2. Anderson, C., "Leveraging Data to Drive Innovation," *Wall Street Journal* Available: https://www.wsj.com/articles/SB10001424127887323468604578245540627666664.

[3] The earliest known recorded mention of "semi-structured data" occurred in 1992 from N. J. Belkin and Croft, W. B., "Information filtering and information retrieval: two sides of the same coin?," in *Communications of the ACM: Special Issue on Information Filtering*, vol. 35(12), pp. 29 – 38. The next two mentions were in 1995 from D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman and J. Widom, "Querying semistructured heterogeneous information," presented at <i>Deductive and Object-Oriented Databases (DOOD '95), LNCS</i>, No. 1013, pp. 319-344, Springer, and M. Tresch, N. Palmer, and A. Luniewski, "Type classification of semi-structured data," in *Proceedings of the International Conference on Very Large Data Bases (VLDB)*. However, the real popularization of the term "semi-structured data" occurred through the seminal 1997 papers from S. Abiteboul, "Querying semi-structured data," in *International Conference on Data Base Theory (ICDT)*, pp. 1-18, Delphi, Greece, 1997 (http://dbpubs.stanford.edu:8090/pub/1996-19) and P. Buneman, "Semistructured data," in *ACM Symposium on Principles of Database Systems (PODS)*, pp. 117-121, Tucson, Arizona, May 1997 (http://db.cis.upenn.edu/DL/97/Tutorial-Peter/tutorial-semi-pods.ps.gz). Of course, semi-structured data had existed before these early references; only it had not been named as such.

[4] Schäuble, P., "SPIDER: A Multiuser Information Retrieval System for Semistructured and Dynamic Data," *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1993, pp. 318–327.

[5] Magnani, M., and Montesi, D., *A Unified Approach to Structured, Semistructured and Unstructured Data*, 2004.

[6] As of December 2017, the count was about 1450, with certainly some formats remaining unlisted. See https://en.wikipedia.org/wiki/List_of_file_formats.

[7] Madhavan, J., Jeffery, S. R., Cohen, S., Dong, X., Ko, D., Yu, C., and Halevy, A., "Web-Scale Data Integration: You Can Only Afford to Pay as You Go," CIDR, 2007.

[8] Singhal, A., "Introducing the Knowledge Graph: Things, Not Strings," *Official Google Blog*, May 2012.

[9] Pluempitiwiriyawej, C., and Hammer, J., *A Classification Scheme for Semantic and Schematic Heterogeneities in XML Data Sources*, Gainesville: FL, 36, 2000.

[10] Robert Hillard notes he agrees with the importance of semantics and graph structures, but also believes the complexity of real-world information and knowledge graphs are major obstacles to the navigation of content.

[11] Plato, "Phaedrus Dialog (page 265e)," *Perseus Digital Library* Available: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0174%3Atext%3DPhaedrus%3Apage%3D265.

[12] Steiner, P., "CS Peirce and Artificial Intelligence: Historical Heritage and (new) Theoretical Stakes," *Philosophy and Theory of Artificial Intelligence*, Springer, 2013, pp. 265–276.

[13] Ayers, M. R., "Locke versus Aristotle on Natural Kinds.," *The Journal of Philosophy*, 1981, pp. 247–272.

[14] Hulswit, M., "Peirce's Teleological Approach to Natural Classes," *Transactions of the Charles S. Peirce Society*, 1997, pp. 722–772.

[15] Bergman, M. K., "Shaping Wikipedia into a Computable Knowledge Base," *AI3:::Adaptive Information*, Mar. 2015.

[16] Courtesy of and adapted from https://commons.wikimedia.org/wiki/File:British_Isles_Euler_diagram_15.svg.