WHAT IS STRUCTURE?

Michael K. Bergman[1], Coralville, Iowa USA
May 28, 2012
AI3:::Adaptive Information blog

One of the main reasons I am such a big fan of RDF as a canonical data model is its ability to capture information in structured, semi-structured and unstructured form [1]. These sources are conventionally defined as:

- *Structured data* — information presented according to a defined data model, often found in relational databases or other forms of tabular data
- *Semi-structured data* — does not conform with the formal structure of data models, but contains tags or other markers to denote fields within the content. Markup languages embedded in text are a common form of such sources
- *Unstructured data* — information content, generally oriented to text, that lacks an explicit data model or schema; structured information can be obtained from it via data mining or information extraction.

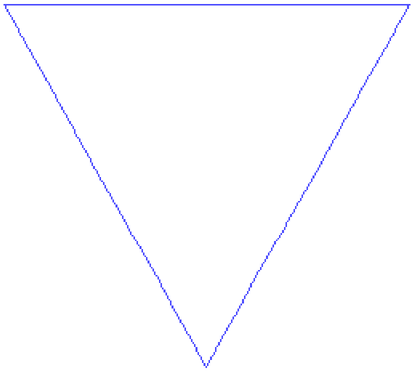
A major trend I have written about for some time is the emergence of the *structured Web*: that is, the exposing of structure from these varied sources in order for more information to be interconnected and made interoperable. I have posited — really a view shared by many — that the structured Web is an intermediate point in the evolution of the Web from one of documents to one where meaningful semantics occurs [2].

It is clear in my writings — indeed in the very name of my company, Structured Dynamics — that structure plays a major role in our thinking. The use and reliance on this term, though, begs the question: just what ***is*** structure in an informational sense? We'll find it helpful to get at the question of *What is structure?* from a basis using first principles. And this, in turn, may also provide insight into how structure and information are in fact inextricably entwined.

A General Definition of Structure

According to Wikipedia, structure is a fundamental notion, of tangible or intangible character, that refers to the recognition, observation, nature, or permanence of patterns and relationships of entities. The concept may refer to an object, such as a built structure, or an attribute, such as the structure of society.

Structure may be abstract, or it may be concrete. Its realm ranges from the physical to ideas and concepts. As a term, "structure" seems to be ubiquitous to every domain. Structure may be found across every conceivable scale, from the most minute and minuscule to the cosmic. Even realms without any physical aspect at all — such as ideas and languages and beliefs — are perceived by many to have structure. We apply the term to any circumstance in which things are arranged or connected to one another, as a means to describe the organization or



[1] Email: mike@mkbergman.com

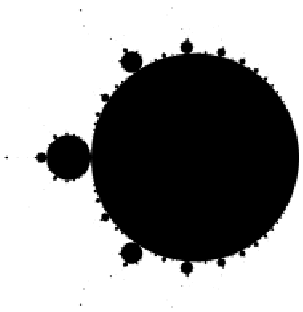
relationships of things. We seem to know structure when we see it, and to be able to discern structure of very many kinds against unstructured or random backgrounds.

In this way structure quite resembles patterns, perhaps could even be used synonymously with that term. Other closely related concepts include order, organization, design or form. When expressed, structure, particularly that of a recognizably ordered or symmetrical nature, is often called beautiful.

One aspect of structure, I think, that provides the key to its roles and importance is that it can be expressed in shortened form as a mathematical statement. One could even be so bold as to say that mathematics is the language of structure. This observation is one of the threads that will help us tie structure to information.

The Patterned Basis of Nature

The natural world is replete with structure. Patterns in nature are regularities of visual form found in the natural world. Each such pattern can be modeled mathematically. Typical mathematical forms in nature include fractals, spirals, flows, waves, lattices, arrays, Golden ratios, tilings, Fibonacci sequences, and power laws. We see them in such structures as clouds, trees, leaves, river networks, fault lines, mountain ranges, craters, animal spots and stripes, shells, lightning bolts, coastlines, flowers, fruits, skeletons, cracks, growth rings, heartbeats and rates, earthquakes, veining, snow flakes, crystals, blood and pulmonary vessels, ocean waves, turbulence, bee hives, dunes and DNA.



The mathematical expression of structures in nature is frequently repeated or recursive in nature, often in a self-organizing manner. The swirls of a snail’s shell reflect a Fibonacci sequence, while natural landscapes or lifeforms often have a fractal aspect (as expressed by some of the figures in this article). Fractals are typically self-similar patterns, generally involving some fractional or ratioed formula that is recursively applied. Another way to define it is as a detailed pattern repeating itself.

Even though these patterns can often be expressed simply and mathematically, and they often repeat themselves, their starting conditions can lead to tremendous variability and a lack of predictability. This makes them chaotic, as studied under chaos theory, though their patterns are often discernible.

While we certainly see randomness in statistics, quantum physics and Brownian motion, it is also striking how what gives nature its beauty is structure. As a force separate and apart from the random, there appears to be something in structure that guides the expression of what is nature and what is so pleasing to behold. Self-similar and repeated structures across the widest variety of spatial scales seems to be an abiding aspect of nature.

Structure in Language

Such forms of repeated patterns or structure are also inherent in that most unique of human capabilities, language. As a symbolic species [3], we first used symbols as a way to represent the ideas of things. Simple markings, drawings and ideograms grew into more complicated structures such as alphabets and languages. The languages themselves came to embrace still further structure via sentence structures, document structures, and structures for organizing and categorizing multiple documents. In fact, one of the most popular aspects of this blog site is its *Timeline of Information History* — worth your look — that shows the progression of structure in information throughout human history.

Grammar is often understood as the rules or structure that governs language. It is composed of syntax, including punctuation, traditionally understood as the sentence structure of languages, and morphology, which is the structural understanding of a language’s linguistic units, such as words, affixes, parts of speech, intonation or context. There is a whole field of linguistic typology that studies and classifies languages according to their structural features. But grammar is hardly the limit to language structure.

Semantics, the meaning of language, used to be held separate from grammar or structure. But via the advent of the thesaurus, and then linguistic databases such as WordNet and more recently concept graphs that relate words and terms into connected understandings, we also have now come to understand that semantics also has structure. Indeed, these very structural aspects are now opening up to us techniques and methods — generally classified under the heading of natural language processing (NLP) — for extracting meaningful structure from the very basis of written or spoken language.

It is the marriage of the computer with language that is illuminating these understandings of structure in language. And that opening, in turn, is enabling us to capture and model the basis of human language discourse in ways that can be codified, characterized, shared and analyzed. Machine learning and processing is now enabling us to complete the virtual circle of language. From its roots in symbols, we are now able to extract and understand those very same symbols in order to derive information and knowledge from our daily discourse. We are doing this by gleaning the structure of language, which in turn enables us to relate it to all other forms of structured information.

Common Threads Via Patterns

The continuation of structure from nature to language extends across all aspects of human endeavor. I remember excitedly describing to a colleague more than a decade ago what likely is a pedestrian observation: pattern matching is a common task in many fields. (I had observed that pattern matching in very different forms was a standard practice in most areas of industry and commerce.) My "insight" was that this commonality was not widely understood, which meant that pattern matching techniques in one field were not often exploited or seen as transferable to other domains.

In computer science, pattern matching is the act of checking some sequence of tokens for the presence of the constituents of some pattern. It is closely related to the idea of pattern recognition, which is the characterization of some discernible and repeated sequence. These techniques, as noted, are widely applied, with each field tending to have its own favorite algorithms. Common applications that one sees for such pattern-based calculations include communications [4], encoding and coding theory, file compression, data compression, machine learning, video compression, mathematics (including engineering and signal processing via such techniques as Fourier transforms), cryptography, NLP [5], speech recognition, image recognition, OCR, image analysis, search, sound cleaning (that is, error detection, such as Dolby) and gene sequence searching and alignment, among many others.

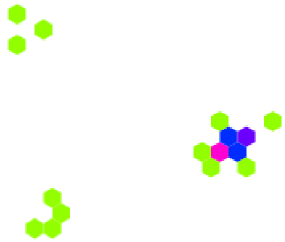
To better understand what is happening here and the commonalities, let's look at the idea of compression. Data compression is valuable for transmitting any form of content in wired or wireless manners because we can transmit the same (or closely similar) message faster and with less bandwidth [6]. There are both lossless (no loss of information) and lossy compression methods. Lossless data compression algorithms usually exploit statistical redundancy — that is, a pattern match — to represent data more concisely without losing information. Redundancy in information theory is the number of bits used to transmit a message minus the number of bits of actual information in the message. Lossless compression is possible because most real-world data has statistical redundancy. In lossy data compression, some loss of information is acceptable by dropping detail from the data to save storage space. These methods are guided by research that indicates, say, how certain frequencies may not be heard or seen by people and can be removed from the source data.

On a different level, there is a close connection between machine learning and compression: a system that predicts the posterior probabilities of a sequence given its entire history can be used for optimal data compression (by using arithmetic coding on the output distribution), while an optimal compressor can be used for prediction (by finding the symbol that compresses best, given the previous history). This equivalence has been used as justification for data compression as a benchmark for “general intelligence.” On a still different level, one major part of cryptography is the exact opposite of these objectives: constructing messages that pattern matching fails against or is extremely costly or time-consuming to analyze.

When one stands back from any observable phenomena — be it natural or human communications — we can see that the “information” that is being conveyed often has patterns, recursion or other structure that enables it to be represented more simply and compactly in mathematical form. This brings me back to my two favorite protagonists in my recent writings — Claude Shannon and Charles S. Peirce.

Information is Structure

Claude Shannon‘s seminal work in 1948 on information theory dealt with the amount of information that could be theoretically and *predictably* communicated between a sender and a receiver [7] [8]. No context or semantics were implied in this communication, only the amount of information (for which Shannon introduced the term “bits”) and what might be subject to losses (or uncertainty in the accurate communication of the message). In this regard, what Shannon called “information” is what we would best term “data” in today’s parlance.



The context of Shannon’s paper and work by others preceding him was to understand information losses in communication systems or networks. Much of the impetus for this came about because of issues in wartime communications and early ciphers and cryptography and the emerging advent of digital computers. But the insights from Shannon’s paper also relate closely to the issues of data patterns and data compression.

A key measure of Shannon’s theory is what he referred to as *information entropy*, which is usually expressed by the average number of bits needed to store or communicate one symbol in a message. Entropy quantifies the uncertainty involved in predicting the value of a random variable. The Shannon entropy measure is actually a measure of the uncertainty based on the communication (transmittal) between a sender and a receiver; the actual information that gets transmitted and predictably received was formulated by Shannon as *R*, which can never be zero because all communication systems have losses.

A simple intuition can show how this formulation relates to patterns or data compression. Let’s take a message of completely random digits. In order to accurately communicate that message, all digits (bits) would have to be transmitted in their original state and form. Absolutely no compression of this message is possible. If, however, there are patterns within the message (which, of course, now ceases to make the message random), these can be represented algorithmically in shortened form, so that we only need communicate the algorithm and not the full bits of the original message. If this “compression” algorithm can then be used to reconstruct the bit stream of the original message, the data compression method is deemed to be lossless. The algorithm so derived is also the expression of the pattern that enabled us to compress the message in the first place (such as `a*2+1`).

We can apply this same type of intuition to human language. In order to improve communication efficiency, the most common words (*e.g.*, “a”, “the”, “I”) should be shorter than less common words (*e.g.*, “disambiguation”, “semantics”, “ontology”), so that sentences will not be too long. As they are. This is an equivalent principal to data compression. In fact, such repeats and patterns apply to the natural world as well.

Shannon’s idea of information entropy has come to inform the even broader subject of entropy in physics and the 2nd Law of Thermodynamics [10]. According to Koelman, “the entropy of a physical system is the minimum number of bits you need to fully describe the detailed state of the system.” Very random (uncertain) states have high entropy, patterned ones low entropy. As I noted recently, in open systems, structures (patterns) are a means to speed the tendency to equilibrate across energy gradients [8]. This observation helps provide insight into structure in natural systems, and why life and human communications tend toward less randomness. Structure will always continue to emerge because it is adaptive to speed the deltas across these gradients; structure provides the fundamental commonality between biological information (life) and human information.

In the words of Thomas Schneider [11], "Information is always a measure of the decrease of uncertainty at a receiver." Of course, in Shannon's context, what is actually being measured here is data (or bits), not information embodying any semantic meaning or context. Thus, the terminology may not be accurate for discussing "information" in a contemporary sense. But it does show that "structure" — that is, the basis for shortening the length of a message while still retaining its accuracy — is information (in the Shannon context). In this information there is order or patterns, often of a hierarchical or fractal or graph nature. Any structure that emerges that is better able to reduce the energy gradient faster will be favored according to the 2nd Law.

Still More Structure Makes "Information" Information

The data that constitutes "information" in the Shannon sense still lacks context and meaning. In communications terms, it is data; it has not yet met the threshold of actionable information. It is in this next step that we can look to Charles Sanders Peirce (1839 – 1914) for guidance [9].

The core of Peirce's world view is based in semiotics, the study and logic of signs. In his seminal writing on this, "What is in a Sign?" [10], he wrote that "every intellectual operation involves a triad of symbols" and "all reasoning is an interpretation of signs of some kind". A sign of an object leads to interpretants, which, as signs, then lead to further interpretants. Peirce's triadic logic of signs in fact is a taxonomy of sign relations, in which signs get reified and expanded via still further signs, ultimately leading to communication, understanding and an approximation of "canonical" truth. Peirce saw the scientific method as itself an ultimate example of this process. The key aspect of signs for Peirce is the ongoing process of interpretation and reference to further signs.

Ideograms leading to characters, that get combined into sentences and other syntax, and then get embedded within contexts of shared meanings show how these structures compound themselves and lead to clearer understandings (that is, accurate messages) in the context of human languages. While the Shannon understanding of "information" lacked context and meaning, we can see how still higher-order structures may be imposed through these reifications of symbols and signs that improve the accuracy and efficiency of our messages. Though Peirce did not couch his theory of semiosis on structure nor information, we can see it as a natural extension of the same structural premises in Shannon's formulation.

Information is structure, and structure is information.

In fact, today, we now see the "structure" in the semantic relationships of language through the graph structures of ontologies and linguistic databases such as WordNet. The understanding and explication of these structures are having a similarly beneficial effect on how still more refined and contextual messages can be composed, transmitted and received. Human-to-machine communications is (merely!) the challenge of codifying and making explicit the implicit structures in our languages.

The Peirceian ideas of interpretation (context) and compounding and reifying structure are a major intellectual breakthrough for extending the Shannon "information" theory to information in the modern sense. These insights also occur within a testable logic for how things and the names of things can be understood and related to one another, via logical statements or structures. These, in turn, can be symbolized and formalized into logical constructs that can capture the structure of natural language as well as more structured data (or even nature, as some of the earlier Peirce speculation asserts [13]).

According to this interpretation of Peirce, the nature of information is the process of communicating a form from the object to the interpretant through the sign [14]. The clarity of Peirce's logic of signs is an underlying factor, I believe, for why we are finally seeing our way clear to how to capture, represent and relate information from a diversity of sources and viewpoints that is defensible and interoperable.

Structure is Information

Common to all of these perspectives — from patterns to nature and on to life and then animal and human communications — we see that structure is information. Human artifacts and technology, though not "messages" in a conventional sense, also embody the information of how they are built within their structures [15]. We also see the interplay of patterns and information in many processes of the natural world [16] from complexity theory, to emergence, to autopoiesis, and on to autocatalysis, self-organization, stratification and cellular automata [17]. Structure in its many guises is ubiquitous.

We, as beings who can symbolically record our perceptions, seem to innately recognize patterns. We see beauty in symmetry. Bilateral symmetry seems to be deeply ingrained in the inherent perception by humans of the likely health or fitness of other living creatures. We see beauty in the patterned, repeated variability of nature. We see beauty in the clever turn of phrase, or in music, or art, or the expressiveness of a mathematical formulation.

We also seem to recognize beauty in the simple. Seemingly complex bit streams that can be reduced to the short, algorithmic expression are always viewed as more elegant than lengthier, more complex alternatives. The simple laws of motion and Newtonian physics fit this pattern, as does Einstein's $E=mc^2$. This preference for the simple is a preference for the greater adaptiveness of the shorter, more universal pattern to messages, an insight indicated by Shannon's information theory.

In the more prosaic terms of my vocation in the Web and information technology, these insights point to the importance of finding and deriving structured representations of information — including *meaning* (semantics) — that can be simply expressed and efficiently conveyed. Building upon the accretions of structure in human and computer languages, the semantic Web and semantic technologies offer just such a prospect. These insights provide a guidepost for how and where to look for the next structural innovations. We find them in the algorithms of nature and language, and in making connections that provide the basis for still more structure and patterned commonalities.

Ideas and algorithms around loseless compression and graph theory and network analysis are, I believe, the next fruitful hunting grounds for finding still higher-orders of structure, which can be simply expressed. The patterns of nature, which have emerged incrementally and randomly over the eons of cosmological time, look to be an excellent laboratory.

So, as we see across examples from nature and life to language and all manner of communications, information is structure and structure is information. And it is simply beautiful.

Acknowledgements

This article was originally posted on the *AI3:::Adaptive Information* Web site at http://www.mkbergman.com/1011/what-is-structure/. This version has been edited and reformatted slightly for PDF distribution. We thank Cognonto Corporation for making this content freely available.

[1] I discuss this advantage, among others, in M. K. Bergman, 2009. "Advantages and Myths of RDF," *AI3:::Adaptive Innovation* blog, April 8, 2009. See http://www.mkbergman.com/483/advantages-and-myths-of-rdf/.

[2] The *structured Web* is object-level data within Internet documents and databases that can be extracted, converted from available forms, represented in standard ways, shared, re-purposed, combined, viewed, analyzed and qualified without respect to originating form or

provenance. See further M. K. Bergman, 2007. “What is the Structured Web?,” *AI3:::Adaptive Innovation* blog, July 18, 2007. See http://www.mkbergman.com/390/what-is-the-structured-web/. Also, for a diagram of the evolution of the Web, see M. K. Bergman, 2007. “More Structure, More Terminology and (hopefully) More Clarity,” *AI3:::Adaptive Innovation* blog, July 22, 2007. See http://www.mkbergman.com/391/more-structure-more-terminology-and-hopefully-more-clarity/.

[3] Terrence W. Deacon, 1997. *The Symbolic Species: The Co-Evolution of Language and the Brain*, W. W. Norton & Company, July 1997 527 pp. (ISBN-10: 0393038386)

[4] Communications is a particularly rich domain with techniques such as the Viterbi algorithm , which has found universal application in decoding the convolutional codes used in both CDMA and GSM digital cellular, dial-up modems, satellite, deep-space communications, and 802.11 wireless LANs.

[5] Notable areas in natural language processing (NLP) that rely on pattern-based algorithms include classification, clustering, summarization, disambiguation, information extraction and machine translation.

[6] To see some of the associated compression algorithms, there is a massive list of “codecs” (compression/decompression) techniques available; fractal compression is one.

[7] 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.

[8] I first raised the relation of Shannon’s paper to data patterns — but did not discuss it further awaiting this current article — in M. K. Bergman, 2012. “The Trouble with Memes,” *AI3:::Adaptive Innovation* blog, April 4, 2012. See http://www.mkbergman.com/1004/the-trouble-with-memes/.

[9] I first overviewed Peirce’s relation to information messaging in M. K. Bergman, 2012. “Give Me a Sign: What Do Things Mean on the Semantic Web?,” *AI3:::Adaptive Innovation* blog, January 24, 2012. See http://www.mkbergman.com/994/give-me-a-sign-what-do-things-mean-on-the-semantic-web/. Peirce had a predilection for expressing his ideas in “threes” throughout his writings.

[10] For a very attainable lay description, see Johannes Koelman, 2012. “What Is Entropy?,” in Science 2.0 blog, May 5, 2012. See http://www.science20.com/hammock_physicist/what_entropy-89730.

[11] See Thomas D. Schneider, 2012. “Information Is Not Entropy, Information Is Not Uncertainty!,” Web page retrieved April 4, 2012; see http://www.lecb.ncifcrf.gov/~toms/information.is.not.uncertainty.html.

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

[13] It is somewhat amazing, more than a half century before Shannon, that Peirce himself considered “quasi-minds” such as crystals and bees to be sufficient as the interpreters of signs. See *Commens Dictionary of Peirce’s Terms* (CDPT), Peirce’s own definitions, and the entry on “quasi-minds”; see http://www.helsinki.fi/science/commens/terms/quasimind.html.

[14] João Queiroz, Claus Emmeche and Charbel Niño El-Hania, 2005. “Information and Semiosis in Living Systems: A Semiotic Approach,” in SEED 2005, Vol. 1. pp 60-90; see http://www.library.utoronto.ca/see/SEED/Vol5-1/Queiroz_Emmeche_El-Hani.pdf.

[15] Kevin Kelley has written most broadly about this in his book, *What Technology Wants*, Viking/Penguin, October 2010. For a brief introduction, see Kevin Kelly, 2006. “The Seventh Kingdom,” in The Technium blog, February 1, 2006. See http://www.kk.org/thetechnium/archives/2006/02/the_seventh_kin.php.

[16] For a broad overview, see John Cleveland, 1994. “Complexity Theory: Basic Concepts and Application to Systems Thinking,” *Innovation Network for Communities*, March 27, 1994. May be obtained at http://www.slideshare.net/johncleveland/complexity-theory-basic-concepts.

[17] For more on cellular automata, see Stephen Wolfram, 2002. *A New Kind of Science*, Wolfram Media, Inc., May 14, 2002, 1197 pp. ISBN 1-57955-008-8.