Untangling the Web of Technical Knowledge

A Model of Information Content and Structure

By
David Wojick, Ph.D.
dwojick@hughes.net

For
Office of Scientific and Technical Information
U.S. Department of Energy
www.osti.gov

October, 2010

I. Introduction

II. What is information content?

III. The three basic elements of information content

IV. Information structures based on the three elements

V. The case of information system design

VI. Visualization of information structures

VII. Historical note

I. Introduction.

This is the information age, but what is information? Several years ago, while working on an information system design problem, I realized that there is no scientific definition of "information content", as that term is ordinarily used. Traditional information theory does not provide one; because information theory does not take into account that information has content. It accepts random strings of letters as information.

So I set out to do what in mathematical logic is called a "rational reconstruction". This means to develop a formal scientific definition of an ordinary language concept. In fact, I have found not one, but 3 definitions of information content, nested one within another. These 3 definitions turn out to be quite powerful. They provide the basis for a new model of information content and structure, one that is applicable to any information design activity. In fact, the model supports several new design methods, including actually measuring and optimizing content in various interesting ways.

In my vision people cruise through the structure of their information as one would fly a starship. They visit strange and beautiful worlds, worlds that are in a sense already real. We just can't see them yet.

Some aspects of this model are merely conjectural at present. A research program will be required to test and refine its many facets. However, because the model grows out of 37 years of study and visualization of the structure of complex situations, some of its applications have already been demonstrated.

II. What is information content?

The 3 proposed definitions, or types, of information content are:

Type 1: The propositional content of expressed thought

Type 2: The propositional content of symbolic expression or display

Type 3: The propositional content of meaningful expression or display

In all three types the information content is propositional in nature. Propositions are the fundamental objects studied in mathematical logic. At its simplest, a proposition is the meaning of a sentence independent of the language used to express that meaning. You can say that snow is white in many languages, so there is something --the proposition that snow is white -- that is independent of the language you use.

While unfamiliar to most people, this technical concept of a proposition is explained in most logic texts. In fact, many practitioners of mathematical logic consider that discipline to be the science of propositions. So if these definitions are accepted as capturing the ordinary language concept of information content, then we have a solid grounding in a well developed discipline.

The basic difference between the three types of information content is this:

Type 1 information content (expressed thought) requires an expression of thought, normally by either speaking or writing, so it can only be produced by thinking things using language. This is the core definition. Information content is produced when people express propositions by speaking or writing. Written content is of particular importance because it is easily reserved. Most bodies of information consist of this kind of written content.

Type 2 information content (symbolic expression or display) includes Type 1 content. However, it can also be produced by devices as well as by people, but still requires something like language. This captures the fact that we get information from symbol using devices like meters and computers. In the new digital age we get vast quantities of information content from such devices.

Type 3 information content (meaningful expression or display) allows for non-symbolic sources of content such as a simulation or a video. Unlike types 1 and 2, the fact that this type 3 kind of content does not use language makes it a challenge to say what propositions are expressed. However, if a video contains spoken or written content that is type 1, not type 3.

(Caveat: This model of information content has nothing per se to do with issues of the utility of information. It does not distinguish good information from bad, true from false,

important from trivial, used from neglected, etc. Using the model should make it easier to make these distinctions, but they are not specifically addressed. The focus is on the nature and structure of content, not its quality or utility.)

The power of these definitions derives from the fact that expressing a proposition is in a way a very simple act. Indeed, I argue below that any such act can be considered as consisting of just three basic elements. Understanding these elements leads us to a new understanding of information content, as well as to the discovery of a rich world of information structures that underlies all bodies of information.

III. The three basic elements of information content

From the point of view of mathematical logic, any body of expressed propositions is made up of the following three basic elements:

Element 1. The physical context of the expression.

For type 1 this is typically who said or wrote what, including the actual language used, when, where, etc. For type 2 devices this will include actual displays, printouts, etc. For type 3 non-symbolic expressions it is the facts about the video, etc.

In other words, according to our three definitions information is always a tangible thing created by a specific act of expression at a specific time and place. It may be sounds, marks on paper, a dial reading, a video, even an action, but it is always tangible. Thus on this view information content is never an intangible something in someone's head. The latter may be knowledge or belief, but it is not information. We are therefore talking about something that always has a physical aspect.

Note too that information content has to be expressed. Thus simply seeing that a tree is in the road, an act of perception, does not involve information content. This is a significant restriction, because we sometimes use the word information to refer to perception, or even to physical causality. Those uses are excluded from this model.

Element 2. The propositions expressed.

The propositions expressed are the actual, meaningful content of the information. In writing and speaking it is what our sentences say, independent of which language is being used. Propositions are whole thoughts, not phrases or words.

In mathematical logic there are techniques for taking sentences and identifying the propositions expressed. If a piece of type 1 writing or speaking is translated into symbolic logic we can even say precisely how many atomic propositions are expressed. That is, we can measure the information content. Doing this for types 2 or 3 may be more problematic.

Element 3. The things referred to by these propositions.

These things are called in logic the "referents" of the propositions. Reference is discussed in many logic textbooks. Note that referents can be activities, properties, or anything that can be talked about. To say that the snow is white is to refer to snow the stuff (and possibly whiteness the attribute). To say that snow is melting is to refer to snow the stuff and the activity of melting. Referents do not have to be real, nor do the propositions have to be true. Novels and lies have content.

IV. Information structures based on the three elements

As we have seen, any instance of information content, call it a piece of information, involves (1) one or more propositions, (2) expressed in a given physical context, and (3) referring to certain things. This sounds abstract, but in any given case these elements are often pretty obvious. (Or some of them are obvious, while others are not, but I will not go into that issue here.) What is important is that different pieces of information can be fit together according to how their elements are related. This leads to a rich new science that I call information structures.

Given the three basic elements of information content, we note the following:

A. Information content has relational structure. Any two pieces of information content can be related to one another by one of these basic elements, or by a combination of them. The question then becomes: what are the most important relationships that underlie a given body of information? How do the pieces fit together? As it turns out, some of the most important relations are well known, while others are less so.

B. Information content may have several different structures at once. Moreover, there may be a number of very different relationships that are important in a given body of information. If so then when we try to understand that information we are in fact trying to do several things at once, that is, to master several elemental relationships at once. I believe this is one of the chief obstacles to efficient learning. We try to grasp several different relationships without differentiating them.

I call the array of information units that are related by a given relationship an "information structure". Any body of important information includes a number of different, yet significant, structures. Certainly this is true for every body of information I have analyzed so far. In fact many attempts at organizing complex information content suffer from confusing and combining multiple structures.

By using the model you can systematically identify the most important information structures that underlie a given body of information. One can even quantify and measure, as well as display, these structures in various useful ways. Let us consider briefly some of the more common and typically important systems of relations, i.e. structures that obtain among the three basic elements in a body of information.

a. Physical context-based structures.

Alphabetical listing is a common way to relate pieces of information content that is based on their physical form (i.e., the language used). Chronologies of the events that produce information, such as speeches or scientific articles, are another example. So are logs of gauge readings, for that matter.

b. Proposition based structures.

Relations of mathematics and logic are typically proposition based. The former includes spreadsheet information, business or engineering calculations, computer software functional designs, etc. (Of course the referents of these manipulations are also going to be related in important ways, as discussed below). Logic relations include implication and contradiction, both of which are important in computer science, problem solving, the law, etc. Some of these relations are well known. Textbooks tend to be organized around a propositional structure, one which builds new concepts out of prior ones.

Less well known is the system of propositional relations displayed by the issue tree diagram I developed at Carnegie Mellon in the 1970's. I now conjecture that this diagram displays the fundamental relationship between the propositions expressed in most bodies of information. If my definitions are correct the issue tree is the fundamental structure of type 1 information content (writing and speaking).

See http://www.stemed.info/reports/Wojick_Issue_Analysis_txt.pdf and http://www.stemed.info/Repo_Tree.pdf

c. Referent-based structures.

Identical reference (i.e., being about the same thing) is a common relation among some of the pieces of content in a body or system of information. However, it is seldom the case that all of the pieces of information content in a given body of information are related by identical reference.

Rather, the propositions typically refer to various members of one or more systems of related referents. This is because things can be related to one another in so many different and important ways. Great care is necessary to distinguish the different kinds of relations between the referents in a body of information, because they define different information structures. Moreover, in many cases we are ignorant of the important ways that things are related. That's what science is about after all.

For example, it may be that all of the sentences in a cookbook are about cooking. But only some refer to fish, and only some refer to frying. Note that a good index might identify both of these structures. In fact the standard index is tacit recognition that there are other important groupings of content in a book. This is why full text search is so powerful.

d. Hybrid structures.

There are some important relations among pieces of information that are hybrids of the above. For example, we sometimes express propositions that refer to other, previously expressed, propositions or groups of propositions. In logic the former are called “meta level” expressions. Likewise, poetry seems often to depend on relations that combine the audible properties of a physical language with propositional properties.

Ranking of search results is another hybrid case, as it combines consideration of the meaning of sentences with the specific words that occur in them. Note that the standard book index is a precursor to full text search. It is an alphabetical listing of major concepts that are scattered throughout the book.

V. The case of information system design

Our model of information content and structure has major implications for the design of information systems. For example, because of a failure to distinguish between the various important structures that relate information content, many information systems reflect a jumble of partial structures.

At the other extreme are systems designed around a single structure, say a telephone book. This design tends to minimize the availability of the other, often equally important, structures to the user. Kinship or geographical relations of the listed parties, for example. These structures are often available only through laborious analysis. Likewise a cookbook that is organized around kinds of meat, or stages of a meal, will have ways of cooking such as baking scattered throughout.

Given that there may be 10, 20 or 30 important structures in a given body of information content, it follows that any one way of organizing will disorganize all the others. In short, there is no single right way to organize a body of information. Suppose there are ten significant structures underlying a given body (or system) of information content. If the content is organized around just one of these then most users may find that organization unhelpful, as they are looking for one of the other structures. We might call this the dilemma of categorization. Note that it applies to taxonomies and ontologies as well. This dilemma is what makes full text search so powerful. It does not presuppose any given structure.

Information system as I use it here is a very broad term, encompassing such diverse creations as data bases, financial systems, management and executive information systems, interactive courseware, command and control systems, etc. Even books and magazines are information systems. While many system design issues are different for these different sorts of systems, two fundamental issues are always present:

Design issue#1. What information is to be included in the system?

Design issue#2. How should this information be organized?

To vastly oversimplify the matter, using our model these two issues really come down to the question of which structures to incorporate (issue #2), and how much of each (issue #1)? But what would traditionally be the second design issue is in fact now the first. Thus we seem to be designing our information systems in a backward fashion, piling on information content then worrying about its structure, when the structures are already in the content. Moreover, once the significant structures are selected, the question of what information to include becomes largely one of level of detail and unit cost.

VI. Visualization of information structures

Then there is the emerging issue of visualization of information content. Our model is directly related to the issue of visualization of technical information. This is because it introduces a way to look for the hidden structures that underlie every significant body of information. There are basically two things one needs to realize. First is that there may be 10, 20 or even 30 different important structures underlying any significant body of information. These need to be separated for visualization. Second, some of these structures are 3 dimensional, so 2D views are just a mess. It is 3D if you can't draw it without having a lot of crossing lines. What we really need is 3D fly thru visualization of complex information.

I envision the visual navigation of important information structures in virtual reality. Indeed I have already started experimenting with such navigation, using simple 3D navigation software. In my vision people cruise through the structure of their information as one would fly a starship. They visit strange and beautiful worlds, worlds that are in a sense already real. We just can't see them yet.

For example, there are hundreds of thousands of peer reviewed scientific articles published each year. What is the structure of all this? See, for example, http://www.osti.gov/ostiblog/home/entry/sharing_results_is_the_engine and http://www.osti.gov/ostiblog/home/entry/leaping_concepts_and_global_discovery

What is the structure of cancer research? Or energy storage, genomics, etc? Where is the frontier and how is it moving? It seems clear that the answer to questions like this can only be grasped through 3D visualization of the underlying information structures that characterize these fields.

VII. Historical note

The information system referred to earlier, that led to my formulation of this model of information content and structure, was a hypermedia system to support training and the “hot line” for the Standard Army Automated Contracting System. My goal was to develop a taxonomy of hyperlink types, so I asked myself "what are all the ways that two or more pieces of information can be related?" It was at this point that I realized I did not know scientifically what information content was. The model proposed here is the beginning of a precise answer to that question. My sponsor was CACI's Fontaine Moore, who was ahead of her time.

The original concept of information structures began somewhat earlier, with my work for the Office of Naval Research, developing an executive information system for the Chief of Naval Research. I studied the kinds of information the Chief was asking for and began to see that it amounted to various combinations and permutations of a finite number of underlying structures.

The first elements of this model, including the type 1 definition, were first articulated in a submission to the National Institute of Standards, in conjunction with their 1990 conference on issues of standardization in hypertext. At that time I argued for a standard language of linking. Interestingly, among the 40 or so participants in this conference was Tim Berners-Lee, who went on to invent the hypertext World Wide Web shortly thereafter. Whether my model influenced him I have no idea, but the Web has made hypertext one of the world's primary communication media. More recently we have the grand challenge of translating Web content into mobile applications. As a result, the structure of information is now a primary concern.

I also want to thank OSTI director Walt Warnick for requesting this report, as well as for a great deal of encouragement and support.