Chapter 3

Integrative Strategies

The scientific revolution has dramatically expanded our knowledge of the universe while also revealing its even larger, unexplored dimensions. The largest library of ancient times, the Library of Alexandria, held roughly 500,000 books at its peak in the 3rd century BC. In 1992, the Library of Congress reached 100 million books , two hundred times as big (not including magazines, newspapers, digital and film materials) [3-1]. A study by the University of California Berkeley in 2003 estimates that the total world production of print material, including books, newspapers and office documents, is currently around 1.5 exabytes (a billion gigabytes). This is thirty times the current size of the Library of Congress *every year* [3-2].

While population growth and mass print technology have effected large increases in information, a significant factor contributing directly to knowledge growth is the increased specialization of scientific research. As our scientific and conceptual instruments become more precise, we can resolve increasingly detailed knowledge. Thus the need to integrate and synthesize this knowledge becomes increasingly critical. This chapter explores the growth of knowledge and the need for interdisciplinary convergence in knowledge systems.

3.1. Interdisciplinarity

Knowledge organization is an inherently interdisciplinary activity. We must consider the full range of disciplines when there is a desire to organize knowledge content broadly. In addition, the various solutions to knowledge organization itself are also interdisciplinary, requiring computer science, philosophy, visualization and aesthetic design. A general solution to the organization of knowledge should consider the relationship between each of these disciplines.

The most general system would be capable of representing all types of content at a meaningful semantic level. Of course, it is always possible to simply provide an external reference to the raw data of any field. The world wide web permits us to download complete data sets, but the goal here is meaningful representation - we would like to incorporate the actual knowledge into a common framework on a semantic, grammatic level.

Linguistics tells us the semantics of information is greatly determined by its context, which requires us to be explicit about these differences [3-3]. To apply *pressure* in a political sense has a very different meaning from applying *pressure* in mechanical engineering. A general knowledge system must be able to distinguish these meanings and provide for them.

Two disciplines may even conflict, as is the case with theology and biology on the origin of the universe. Both these views are important, so we wish to create systems that are capable of representing multiple belief systems. One implication of this approach is that every piece of knowledge, even seemingly universal ones, are relative to the groups who hold them. We now know the Earth is round, but at one point is was widely believed to be flat. Thus knowledge from that era must be considered in that context, and not our current one. If we understand that all represented knowledge must be associated with an observer - who has a specific background in a specific time - we can construct a system that is capable of supporting changes in human understanding and knowledge that is true in one context while false in others. To the contemporary artist or the poet, the Earth need not be round. This requires an ability to represent experimental, creative views in addition to widely accepted ones without systemic bias.

3.2. Knowledge vs. Information

Up to this point I have used the terms knowledge and information loosely. However, the difference between them is by no means clear even among those engaged in knowledge systems. Dick Stenmark at IT University of Gothenburg states:

> "It has often been pointed out that data, information, and knowledge are not the same, but despite efforts to define them, many researchers use the terms very casually. In particular, the terms knowledge and information are often used interchangeably, even though the two entities are far from identical." [3-4]

Stenmark presents an overview of various working definitions for data, information and knowledge developed by current researchers. The origins of this research can be traced back to the data-information-knowledge-wisdom hierarchy (DIKW) created by Harlan Cleveland and based on a 1934 poem by T.S. Elliot:

> Where is the Life we have lost in living?
>
> Where is the wisdom we have lost in knowledge?
>
> Where is the knowledge we have lost in information?
>
> - T.S. Elliot, 1934

The word information is from the root *inform*, "to give form to the mind". In common usage, information is considered to be the fundamental unit of the internet - everywhere present, moving yet latent.

Data is usually regard to be the least meaningful of the three. However, the meaning content of data is not clearly defined. In one case it is a set of symbols [3-5], in another a set of facts [3-6]. While uninterpreted symbols can be said to carry no meaning (except perhaps in the icon form itself), facts seem to carry the connotation of being "dry" or "static" although they are not at all meaningless.

Information, on the other hand, is thought to be either data with meaning [3-6], data with purpose [3-7], or data flowing as message [3-8]. The last definition is linked to the field of *information theory*, which defines information as a measure of the predictability of a message [3-9]. That these definitions of information vary so widely in this context is one of the challenges of building consistent knowledge systems.

To resolve these ambiguities, the distinctions between *data*, *information* and *knowledge* used in this thesis will be established from the following thought experiment shown in Figure 3.1, called the Astronaut's Experiment.

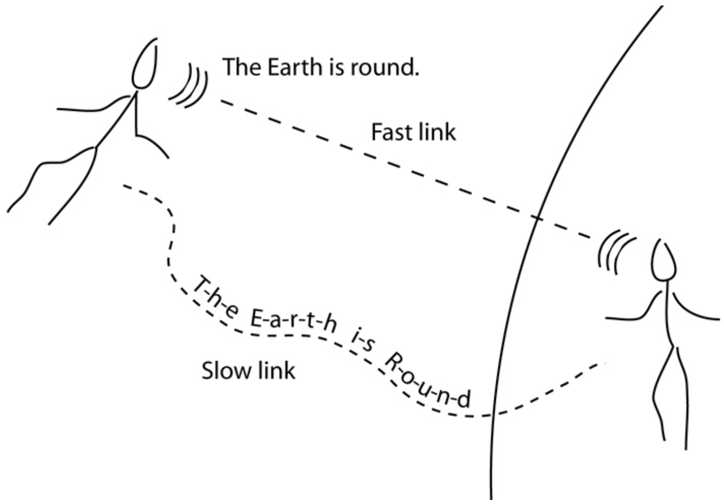

Figure 3.1. The Astronaut's Experiment with
a fast and slow link for communication

The Astronaut's Experiment:

An astronaut is in orbit observing the Earth, while his colleague is on the surface. There are also two radio links, one of which is very slow (several minutes) and text only, the other of which is very fast and allows for voice (a "phone'). The Astronaut sees the Earth and observes its *roundness* directly. He wishes to convey this knowledge to his colleague on the surface. The question is: What is the status of the message while it is being conveyed?

In one case, he simply calls his friend on the phone (fast voice link) and says:

"It can see it, the Earth is round!"

Since the voice is conveyed quickly, the colleague on the ground has an immediate *experience* of that knowledge (although not a direct one). But let us assume instead that the astronaut calls his colleague on the phone (fast link) to say:

I'm sending you some important information.

He then proceeds to use the slow text-only link to send the following message:

T-h-e E-a-r-t-h i-s r-o-u-n-d.

While the message is in transit the listener on the surface does not know what it contains, but the astronaut does. The astronaut does not say, "I am sending you some knowledge" since he cannot place the knowledge directly in the other individual - we have knowledge instantaneously only after we have the experience ourselves. In transit, the message cannot be experienced. When we say "I am sending you some data", it means that I myself do not yet know what it means. But when we say "I am sending you some information" it means that when you read it, it *will* mean something *after* you read it. It is only after we personally comprehend or experience something, through communication with others or by direct experience, that we say "I know that the Earth is round."

The definitions of data, information and knowledge used in this thesis are shown in Table 3.1. It is important not to elevate information to knowledge simply because of our *prior knowledge* (memory). A digital version of Shakespeare residing on disk is known to be meaningful only because our memory. It is our memory that is providing us with a knowledge experience, but this does not imply that a digital version residing on disk has the knowledge *itself* - of which we can only say the "information contains meaning". It is we who contain the actual experience of knowing. Knowledge is both *meaningful* and *experienced* while information is only latently meaningful, and data is neither.

Table 3.1. Definitions of data, information and knowledge with examples.

Data	Raw symbols *not known* to have meaning. Examples: Random numbers, raw data from an experiment not yet analyzed, an intentionally meaningless message sent to another person.
Information	Data present in a system, or in transit, which holds *latent meaning* not yet experienced. Information is data plus the promise that, once read, the data will convey knowledge. Examples: The works of Shakespeare stored on digital media (but not currently being read), a digital 3D model being sent (but not yet received).
Knowledge	The experience of information. Cognition. This is information (latent meaning) after it has been observed and comprehended. Examples: Shakespeare as it exists in me while being read, a personal memory of Shakespeare's writing, a 3D model during visual comprehension.

Several researchers consider knowledge to be *justified or true knowledge* as Plato did, or more loosely *valuable* knowledge [3-4]. However, as the goal of this thesis is to include interdisciplinary and historical ideas it is important to include false knowledge (perhaps previously believed to be true) among these. Therefore knowledge will refer to all experiences regardless of their factuality or truth. True or justified knowledge will be simply referred to as "true knowledge" and "justified knowledge" respectively (subsets of all knowledge).

3.3. Semiotics

The above definitions reveal a critical problem with building knowledge systems, or computational frameworks that "operate on knowledge". Since we have defined knowledge as information-plus-personal-experience we cannot possibly build a machine capable of holding knowledge since the machine is not capable of *experiencing* it. Proponents of strong artificial intelligience (AI) will argue this point, but we need not go so far as to establish machines capable of thought to resolve the issue [3-10].

The emerging field of semasiology or *semiology*, the study of signs and meaning, may help to illuminate this picture. The relationship between words, symbols and ideas goes back as far as Aristotle. This semantic triangle, examined by contemporary linguists such as Ullmann and Baldinger, is normally consider to consisting of 1) the referent (physical object itself), 2) the signifier (fr. *signifiant*), the name or word associated with the object, and 3) the signified (fr. *signifie*), the mental experience or concept [3-11]. The referent is the actual fruit itself (the physical thing), while the word 'apple' is the signifier. The signified is our experience of 'apple' as a concept, and refers to all things which have the quality of being like apples.

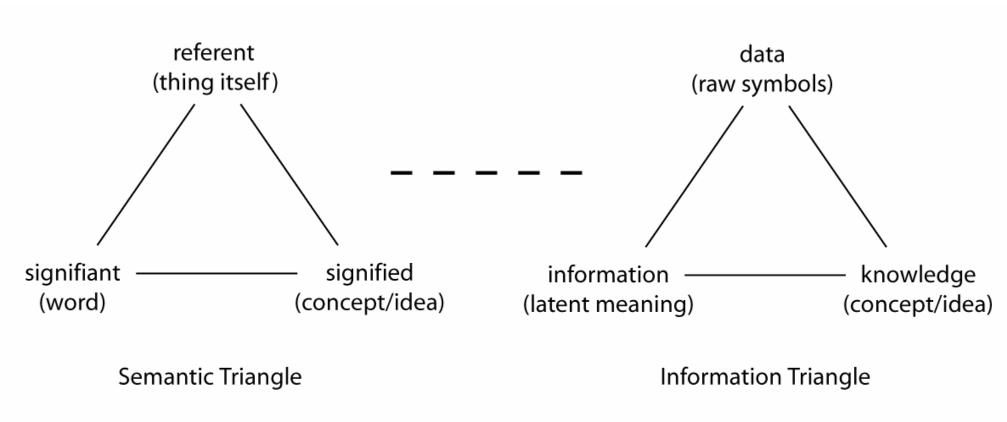

Figure 3.2. Parallels between the Semiotic Triangle and the Information Triangle.

We can observe a unique parallel between data, information and knowledge and the semiotic triangle (Figure 3.2). Data is very much like the *referent*. The former is the raw material of concepts while the latter is the raw material of physical objects. Information is very much like the *signifier* in that a written word is like latent meaning. In fact, a written word *is* latent meaning up until the moment we actual read it. Finally the *signified,* or concept, is very much like knowledge in that the concept encompassing a word is very much like our experience of it. Thus the data-information-knowledge triangle is a conceptual parallel of the referent-signifier-signified triangle.

Interestingly, one of the outcomes of semantic theory is the understanding that in written or spoken language - that is in any representation - the *signified* can never be externally expressed since by writing down an idea it immediately becomes a *signifier*, a word, or information [3-12]. To extend this argument to knowledge and information, the implication is that as soon as knowledge (an internal experience) is written, spoken, or stored it becomes information until it is picked up by someone else. We can never really build *knowledge systems* since by definition computers do not "know", but only manipulate.

Again, proponents of strong AI may argue that computers *can* know, that knowledge experiences are possible by computers, while others may argue that only people experience knowledge and everything else is information or latent knowledge. [1] For the sake of remaining on topic, and to avoid a detailed discussion in metaphysics, I will adopt a simplification.

For practical purposes, I will occasionally refer to *written statements in natural language* as *knowledge* even though they are information according to the definitions above. The justification for accepting natural language as a substitute is that unlike other forms, and because we learn language very early, it requires little or no interpretation to provide a knowledge experience. Reading the statement, "The cat ate my hat." is nearly equivalent to having

[1] To know is to understand, or comprehend. It is possible that machines might achieve this to the same degree as a human being. Then again, perhaps not. More important is the observation that the experience of knowledge is not necessarily a binary activity. There may be degrees in the act of knowing just as there degrees in the scale of biological, chemical and physical activity.

the experience as there is no need to scan through a table of figures or interpret a chart. Language is integral with thinking.

An image may convey perceptual experience more directly, but spoken language is more precise with abstract statements. The statement "The cat ate my hat, which caused me to feel oddly depressed yet amused." is difficult to convey with the image alone. Written natural language thus provides a semantic framework for holding precise abstractions, while images support this with perceptual experience.

To construct a general information system requires a convergence of philosophy, semantic theory, linguistics, aesthetics and information sciences. A summary of the conclusions so far is presented in Table 3.2.

Table 3.2. Summary of definitions for data, information and knowledge.

Concept	**Definition**
Data	Raw, meaningless symbols
Information	Data + Latent meaning
Knowledge (Proper)	Information + Experience (Cognition)
Knowledge (Practical)	Information in Natural and Visual Language (with the goal of future human cognition)
Knowledge System	Systems that operate on Practical Knowledge

3.4. Research as Process

Definiton of knowledge and information in the previous section provide a foundation for a knowledge organization system. The next step is to consider the content of such a system - what we wish to organize. This is the process of constructing an *ontology*, or set of basic categories of existence. As this is a large topic, a more thorough investigation will be considered in chapter six (Ontology & Classification). However, we might reveal integrative solutions by looking more at how we *construct* new knowledge.

While each discipline differs immensely in content the process of academic research provides the context for a common starting point. One of the first activities of an emerging professional in any field is to compile a list of authors, journals and papers in his or her area of study. This forms a foundation for specific interests, a historical framework for study, a source of references for publication, and a list of people doing related research. For the novice in a new field, the first attempt may contain only a few categories (Figure 3.3a).

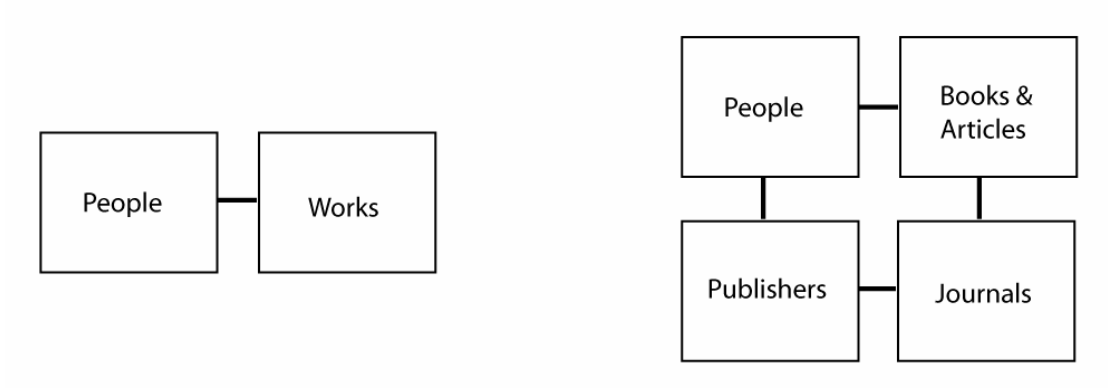

Figure 3.3. Summarizing a field of study. a) First attempt covers only people and works. b) Second attempt includes the means of expression.

While people and works may come first, one quickly realizes the need to list journals and publishers as well (Figure 3.3b). Eventually, with more experience one begins to understand and outline the field itself and to create a taxonomy of the area of study. Examples from music and computer graphics are shown in Table 3.3.

Table 3.3. Taxonomies for the fields of music and computer graphics. (Taxonomies by the author)

Music

I. Composition
   a. Theory
   b. Elements
   c. Technique
II. Performance
   a. Practice
   b. Venue
III. Instruments
   a. Strings
   b. Brass
   c. Woodwind
   d. Percussion
IV. Electronic Music
   a. Recording
   b. Synthesis
   c. Filtering
   d. Hardware
   e. Software
V. Music Production
   a. Producers
   b. Studios
   c. Mastering
VI. Education
   a. Curricula
   b. Resources
   c. Schools
   d. Instructors
VII. History
   a. Movements
   b. Genre
   c. Style

Computer Graphics

I. Mathematics
   a. Linear Algebra
   b. Transformations
   c. Projections
II. Rasterization
   a. Primitives
   b. Text
   c. Curved Surfaces
III. Hidden Surfaces
   a. Depth sorting
   b. Depth culling
   c. Level of detail
IV. Rendering
   a. Local Illumination
   b. Global Illumination
   c. Hybrid Methods
V. Lighting Models
   a. Warn illumination
   b. Phong illumination
VI. Curved Surfaces
   a. Surface smoothness
   b. Representation
   c. Storage
VII. Solid Geometry
   a. Boolean models
   b. Volumetric models
   c. Parametric models
VIII. Animation
   a. Keyframing
   b. Procedural methods
   c. Motion capture

At some point, the researcher may come to understand their chosen discipline well enough that they can enumerate specific ideas, works and relationships (Figure 3.4). For example, the art historian knows the specific works of each artist, the content of those works, and their meaningful relationships. The astrophysicist, due to familiarity, can name all the existing theories for the size and shape of the universe.

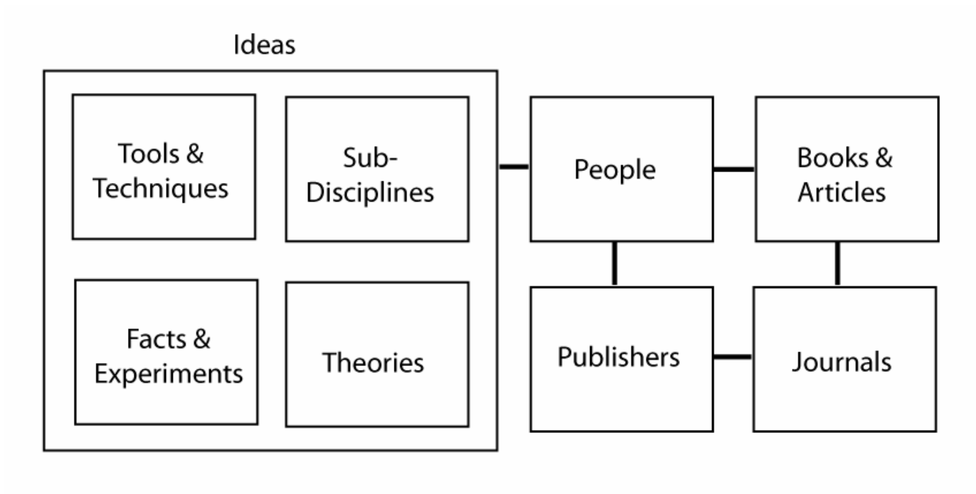

Figure 3.4. Further elaboration of an arbitrary discipline including people, works, publishers, publications and ideas.

What begins as a simple historical database of people and events eventually includes ideas, theorems and relationships. Despite differences in practice or even fundamental differences in belief among various disciplines these patterns of research remain consistent. The practice of making art is very different from mechanical engineering, yet both learn the people and history of their respective fields.

Familiar research materials, such as the MARC format for bibliographic records in digital libraries, consist of the publications and authors of a discipline but usually do not clearly distinguish the ideas present in them. Yet the article-centric nature of these resources, which were created in an era when databases were still being invented, makes it difficult to explore concepts freely.

> "[A bibliographic record] has thus simultaneously (and fortuitously) served an inventory function and a conceptual or informational function. But this is beginning to work less effectively. The increasing incidence of media in new formats has led to a divergence of the two functions, so that records designed for one function do not suffice for the other." [3-14]

In 1959, keywords were introduced to published journal articles to provide a means for machines to search concept rather than just titles and authors [3-15]. Keywords enable search engines but still do not capture the semantic relationships between words. The discussion of natural language processing to resolve this will continue in chapter five (Language & Representation).

A generic approach would incorporate the knowledge of all disciplines simultaneously. Consider the construction of a map of ideas for a number of fields. Many terms, such as that of *energy*, may easily appear in *all* of them. In the physical sciences, that is between chemistry, biology and physics, the concept has one and the same physical meaning but with different interpretations and formulations. In ecology the definition of energy may differ, but the idea must be linked to its more basic physical interpretation to fully appreciate it. In theology and philosophy the idea of *energy* has many other meanings, but these should be linked to the same singular concept as they provide a historical foundation for our modern definitions. Real relationships are lost when concepts, databases, and research areas become distinct. Only by connecting terms across disciplines is it possible to recover this understanding.

A motivating force for this thesis is to construct a general knowledge system to explicitly represent the ideas and relationships of all fields in a common framework. This would allow us to distinguish concepts, organize them, and navigate and explore their meaning freely across the boundaries of current fields of study.