Chapter 2

Background and Context

The ways in which we organize knowledge often seem second nature to us. Immanuel Kant suggests that classification is a fundamental aspect of human nature. While the body of knowledge is vast, the ways in which we organize and communicate knowledge is equally rich. In this chapter, the methods of knowledge organization are divided into three areas: 1) Scientific: the use of systematic methods to classify concepts, 2) Aesthetic: arrangements in which communication and meaning is more important than content, and 3) Computational: methods in which machines assists in organization. These are not necessarily mutually

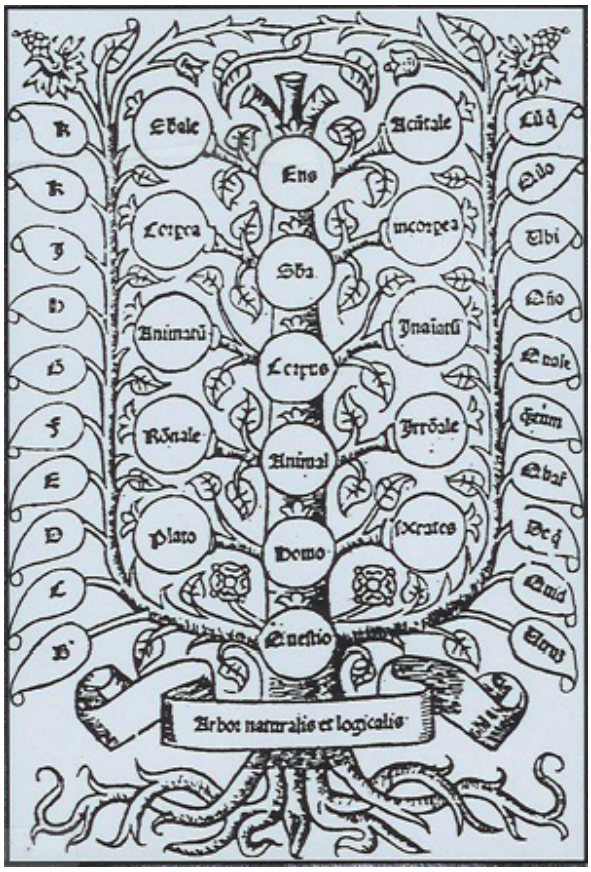


Figure 2.1. Tree of Knowledge, Prophyry, 232 AD

exclusive categories, but for historical reasons it will be easier to examine them distinctly.

2.1 Social and Scientific Knowledge Organization

Organization is deeply embedded in our social structures. Any time there is a need to distinguish objects and communicate these differences we must order what we know. The most common type of organization is the systematic arrangement of physical objects such as the grouping of fruits for sale. Simple groupings are successful until new examples break the pattern. The tomato is scientifically a fruit, but was legally classified as a vegetable by the U.S. Supreme Court in 1893 for taxation reasons:

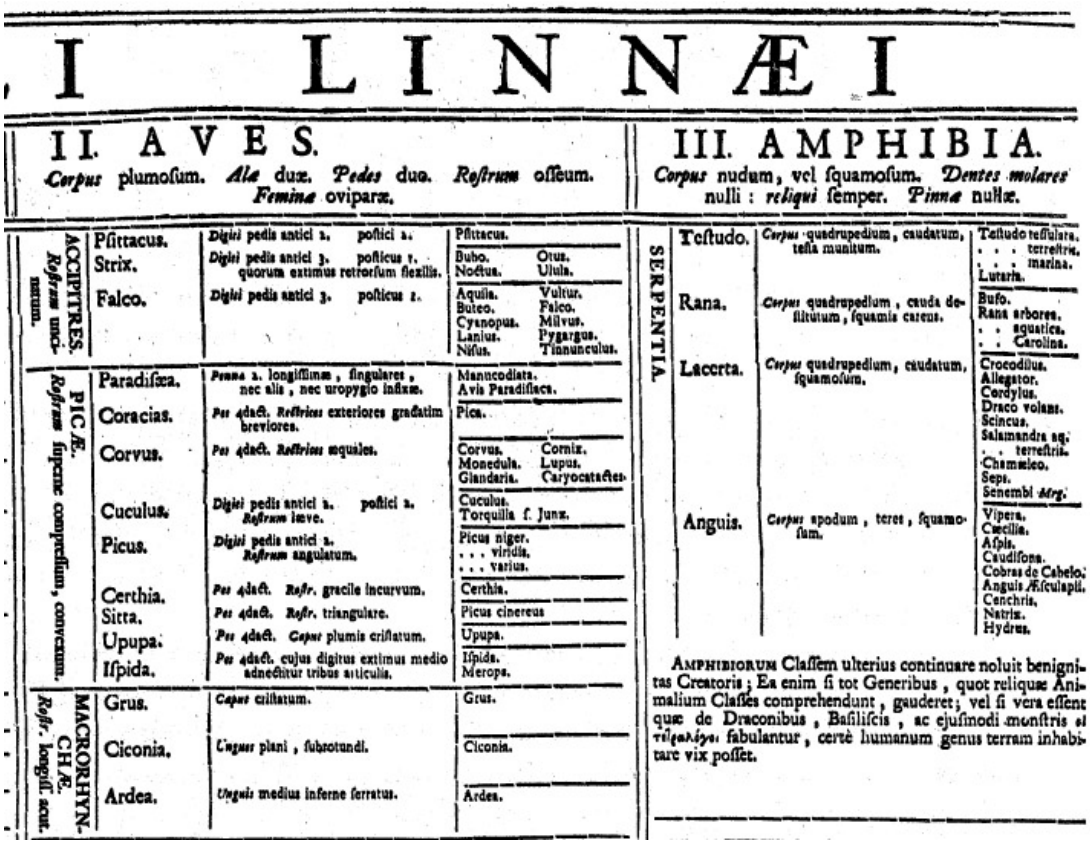
> "Botanically speaking, tomatoes are the fruit of a vine, just as are cucumbers, squashes, beans, and peas. But in the common language of the people, whether sellers or consumers of provisions, all these are vegetables which are grown in kitchen gardens, and which, whether eaten cooked or raw, are, like potatoes, carrots, parsnips, turnips, beets, cauliflower, cabbage, celery, and lettuce, usually served at dinner in, with, or after the soup, fish, or meats which constitute the principal part of the repast, and not, like fruits generally, as dessert." [2-1]

The field of semiotics reveals that we adhere to particular mental arrangements until some new example upsets the balance [2-2]. We are then forced to reorder our concepts. This process is reflected in the sciences in Thomas Kuhn's book *The Structure of Scientific Revolutions* [2-3]. He points

out that in the natural sciences further examination usually reveals subtleties that upset existing theories and structures, requiring new ones.

Before the introduction of systematic classification, concepts were arranged based on observation. For example, in many ancient cultures the elements of fire, water, earth and wind were considered the primary elements of the natural world [2-4]. Prior to the Linnean taxonomy for living things the naturalist John Ray suggested that mad-made classifications are "ultimately arbitrary, and unable to reveal genuine relationships." [2-5]. Systematic classification is the idea that investigation into the true *qualities* of a thing can lead to more correct taxonomies.

The earliest example of diagrammatic knowledge organization can be found in the ancient Greek philosopher Porphyry who sketched the first Tree of Knowledge. It depicts Aristotle's categories of metaphysical being and was the first use of the tree as an organizing principle (Figure 2.1). The term *hierarchy*, meaning sacred rule, was originally used by Pseudo-Dionysius the Areopagite in 1380 to describe three orders of three angels [2-6]. This is one of the earliest examples of a hierarchy with more than one level.



I LINNÆI

II. AVES.

Corpus plumoſum. *Alæ* duæ. *Pedes* duo. *Roſtrum* oſſeum. *Fœmina* oviparæ.

ACCIPITRES. *Roſtrum* uncinatum.	Pſittacus.	*Digiti* pedis antici 2. poſtici 2.	Pſittacus.	
	Strix.	*Digiti* pedis antici 3. poſticus 1. quorum extimus retrorſum flexilis.	Bubo. Noctua.	Otus. Ulula.
	Falco.	*Digiti* pedis antici 3. poſticus 1.	Aquila. Buteo. Cyanopus. Lanius. Niſus.	Vultur. Falco. Milvus. Pygargus. Tinnunculus.
PICÆ. *Roſtrum* ſuperne compreſſum, convexum.	Paradiſæa.	*Penna* 2. longiſſimæ, ſingulares, nec alis, nec uropygio infixæ.	Manucodiata. Avis Paradiſiaca.	
	Coracias.	*Pes* 4dact. *Roſtrum* exteriores gradatim breviores.	Pica.	
	Corvus.	*Pes* 4dact. *Roſtrum* æquales.	Corvus. Monedula. Glandaria.	Cornix. Lupus. Caryocatactes.
	Cuculus.	*Digiti* pedis antici 2. poſtici 2. *Roſtrum* læve.	Cuculus. Torquilla ſ. Junx.	
	Picus.	*Digiti* pedis antici 2. *Roſtrum* angulatum.	Picus niger. . . . viridis. . . . varius.	
	Certhia.	*Pes* 4dact. *Roſtr.* gracile incurvum.	Certhia.	
	Sitta.	*Pes* 4dact. *Roſtr.* triangulare.	Picus cinereus	
	Upupa.	*Pes* 4dact. *Caput* plumis criſtatum.	Upupa.	
	Iſpida.	*Pes* 4dact. cujus digitus extimus medio adnectitur tribus articulis.	Iſpida. Merops.	
MACRORHYNCHÆ. *Roſtr.* longiſſ. acut.	Grus.	*Caput* criſtatum.	Grus.	
	Ciconia.	*Linguæ* plani, ſubrotundi.	Ciconia.	
	Ardea.	*Unguis* medius inferne ſerratus.	Ardea.	

III. AMPHIBIA.

Corpus nudum, vel ſquamoſum. *Dentes molares* nulli: *reliqui* ſemper. *Pinnæ* nullæ.

SERPENTIA.	Teſtudo.	*Corpus* quadrupedium, caudatum, teſta munitum.	Teſtudo teſſulata. . . . terreſtris. . . . marina. Lutaria.
	Rana.	*Corpus* quadrupedium, cauda deſtitutum, ſquamis careus.	Bufo. Rana arborea. . . aquatica. . . Carolina.
	Lacerta.	*Corpus* quadrupedium, caudatum, ſquamoſum.	Crocodilus. Allegator. Cordylus. Draco volans. Scincus. Salamandra aq. . . terreſtris. Chamæleo. Seps. Senembi *Marg.*
	Anguis.	*Corpus* apodum, teres, ſquamoſum.	Vipera. Cæcilia. Aſpis. Caudiſona. Cobras de Cabelo. Anguis Æſculapii. Cenchris. Natrix. Hydrus.

AMPHIBIORUM Claſſem ulterius continuare noluit benignitas Creatoris; Ea enim ſi tot Generibus, quot reliquæ Animalium Claſſes comprehendunt, gauderet; vel ſi vera eſſent quæ de Draconibus, Baſiliſcis, ac ejuſmodi monſtris ex τερατολόγοις fabulantur, certè humanum genus terram inhabitare vix poſſet.

Figure 2.2. Systema Naturae, Carolus Linaeus, 1735 (excerpt)

The most well known modern scientific hierarchy is the taxonomy of living organisms (Figure 2.2). While Aristotle was the first to classify animals based on their means of transport, the modern system of *class*, *order*, *genus* and *species* was developed by Carolus Linnaeus in 1735 in his work *Systema Naturae* [2-7]. In this method of classification, identity follows from an investigation into the *phylogeny* of the organism based on Darwin's *Theory of Evolution by Natural Selection* [2-8]. The tree is effective as an organizing principle in this case because each branch resolves specific differences in the

structure of the organism. Taxonomies are the earliest, but not necessarily ideal method of systematic classification. In chapter six (Ontology and Classification) we investigate how objects with multiple qualities defy placement in a single hierarchy.

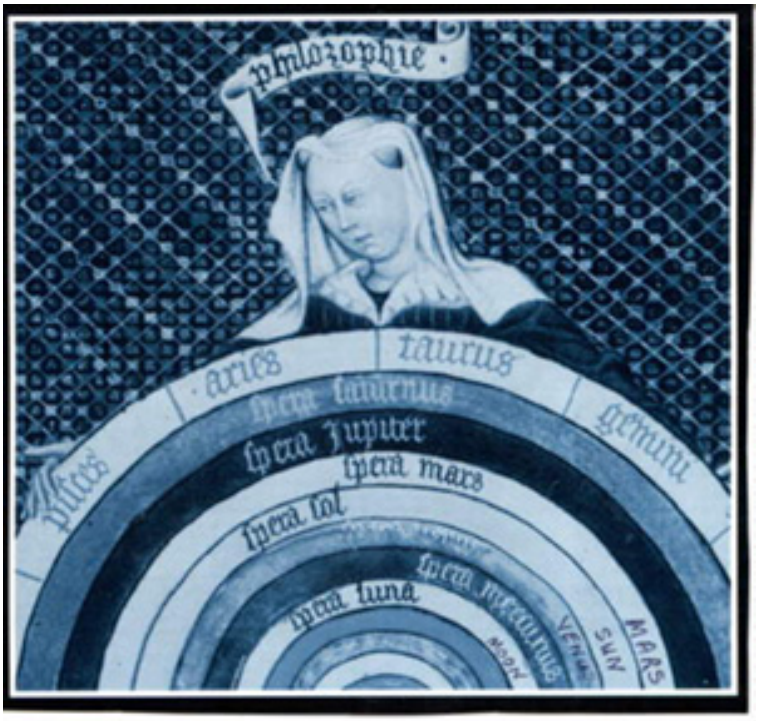


Figure 2.3. Pre-copernican view of the solar system arranged using concentric circles.

The circle can also be viewed as a concept-organizing shape. The ancient Greeks referred to geometry as the perfect description of the heavens [2-9]. In Figure 2.3. we see a pre-copernican view of the solar system arranged on concentric circles. Circles are frequently relied on to depict conceptual relationships and to depict theological, mental and symbolic relationships in

other cultures, as can be observed in the Zodiac, the Mayan calendar and the Tibetan mandala.

We notice that certain types of knowledge, such as astronomy, lend themselves naturally to the geometry of the circle. The structure of an organizational system can therefore be distinguished from its geometry. For example, the concentric system of the planets can be viewed as simply a list, while the geometry of the circle assigns the symbolic or spatial meaning of order from a center.

Although the circle and tree are widely used, they also have certain limitations. Neither is capable of representing knowledge that can be viewed from two *schema,* or methods of classification, simultaneously. In the following example, Table 2.1, a fish can be classified either according to its evolutionary taxa (Linneaus) or according to its means of transport (Aristotle). While the Linnean taxonomy is broadly applicable due to its use of genetic similarity, John Ray was also correct in that the various facets of an object cannot be placed in a singular taxonomy. If we choose a particular hierarchy, we find it difficult study the other quality of a set of objects.

Table 2.1. Classification of fish, bears and human according to two different schema. a) Phylogenic taxonomy, Linneaus and b) Means of transport, Aristotle

Phylogenic taxonomy (Carolus Linneaus)	**Means of transport (Aristotle)**
Animal	Transport
Cordata	Land
Vertebrata	Humans
Fish	Bears
Humans	Sea
Bears	Fish

One solution to the problem of multiple classification introduced by hierarchical arrangement is the network. The term network can be traced to the root *net*, as in a web-like arrangement of threads and wires, while the modern use of network as a means of organization is most apparent in Joseph Novak's formulation of the *concept map* (Figure 2.4).

Novak defines the concept map as follows:

"[Concept maps are] tools for organizing and representing knowledge. They include concepts, usually enclosed in circles or boxes of some type, and relationships between concepts or propositions, indicated by a connecting line between two concepts. Words on the line specify the relationship between the two concepts." [2-10]

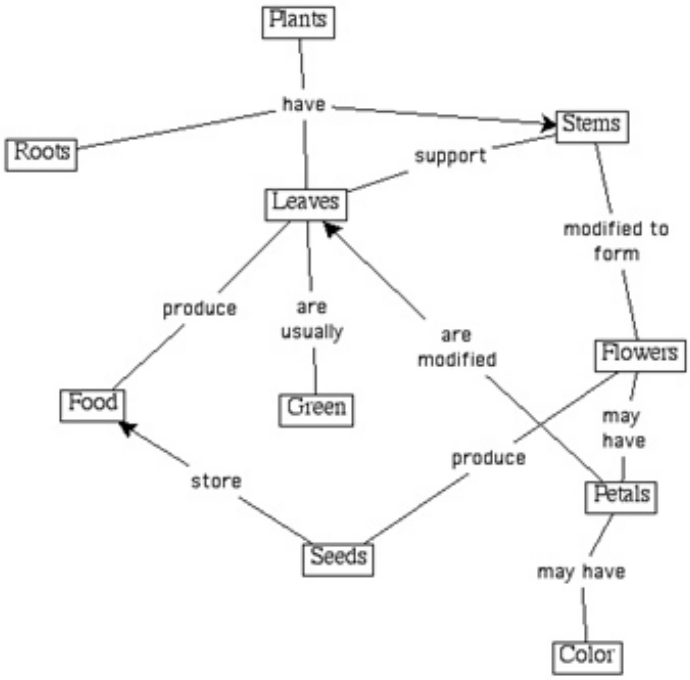


Figure 2.4. Concept map of a plant.

We notice that this definition is fairly unrestricted and consists essentially of a set of concepts (points) and relationships (lines). In discrete mathematics this is called a *graph*, and is the basis of modern computer networks.

Novak originally developed concept maps as an educational tool [2-11]. These were originally constructed by hand. While they are useful when dealing with small numbers of objects, large concepts maps become unwieldy as the amount of information becomes too dense.

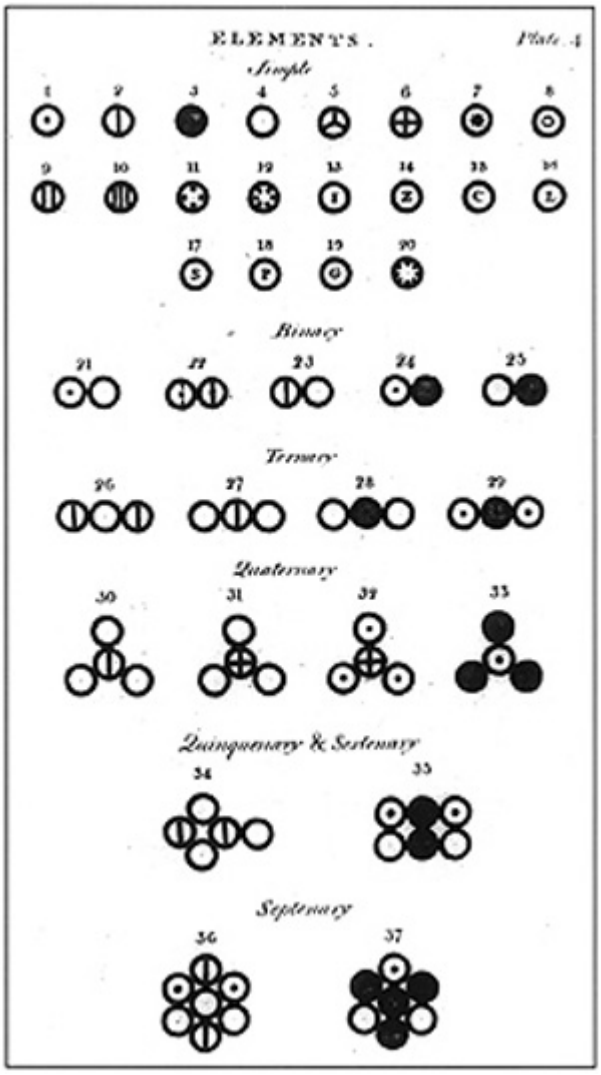


Figure 2.5. John Dalton, New System of Chemical Philosophy, 1808

This issue of visual clutter is discussed further in chapter seven (Visualization). In a sense, concept maps represent the upper limit of the manual expression of systematic knowledge. Even so, concept maps are useful to communicate a set of relational ideas clearly.

Another systematic method of organization is the table. In 1803, John Dalton concluded that chemical interactions must be the result of interactions of atoms of differing *atomic weights.* An image from his 1808 book a *New System of Chemical Philosophy,* Figure 2.5, shows his tabular arrangement of the elements [2-12].

One might suggest that the tabular format is only a visual convenience. As we have seen with the circle, often the geometry of representation can be separated from the structure of its content. In the case of Dalton's table of atoms this is true. However, let us consider the precursor to the modern periodic table of the chemical elements developed by Russian chemist Dmitri Mendeleev.

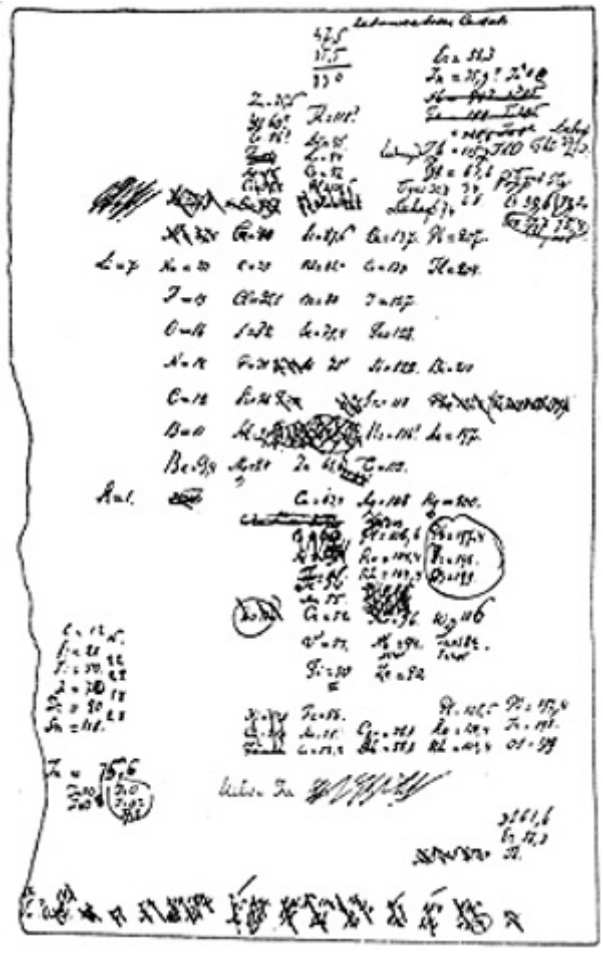


Figure 2.6. Sketches of the periodic table of the elements, Dimitri Mendeleevon, 1869

Mendeleev, in 1869, was also looking at atomic weight. As the electron would not be discovered until 1874, he noticed a certain *periodicity* in the atomic numbers of the elements [2-13]. He thus arranged the atoms in a table, aligning the periodic pattern on each new row.

An early sketch of the periodic table by Mendeleev is shown in Figure 2.6. We see in this case the arrangement is not only for visual communication but also

a matter of necessity. The columns, called groups, represent increasing atomic number while the rows, called periods, were later found to represent the number of free electrons in the atom. In fact, as Mendeleev developed the table it became apparent that certain places must remain blank. He deduced that there must be elements not yet known that belong in these spaces.

The periodic table is by no means the first example of tabular knowledge organization. Mathematicians have used tables to convey numeric relationships for many years. The table need not be rectangular either. Pascal's triangle is a tabular arrangement of numbers in which each is the sum of the numbers found diagonally on the previous rows. The food pyramid is another example of knowledge organization arranged on a triangle.

2.2. The Aesthetics of Arrangement

While the goal of systematic organization is to reflect the world, aesthetic arrangement places emphasis on the need to communicate something more than the relationships between individual elements. Communication by the arrangement of objects is common in the visual arts. Found object art, or *assemblage*, involves the collect of objects to convey a message large than the collection itself.

Figure 2.7. O'Clock, Arman, 1998

In this 1998 work titled O'Clock, Figure 2.7, the French artist Arman refers to his technique as *accumulation*. Unlike Tony Cragg's identical knobs, his selection of clocks emphasizes the wide variety and unique differences between clocks. If we were looking to purchase a clock, we would most likely find them grouped according to type, appearance and features. By arranging these semi-similar objects randomly, Arman directly emphasizes to us their *differences in type.*

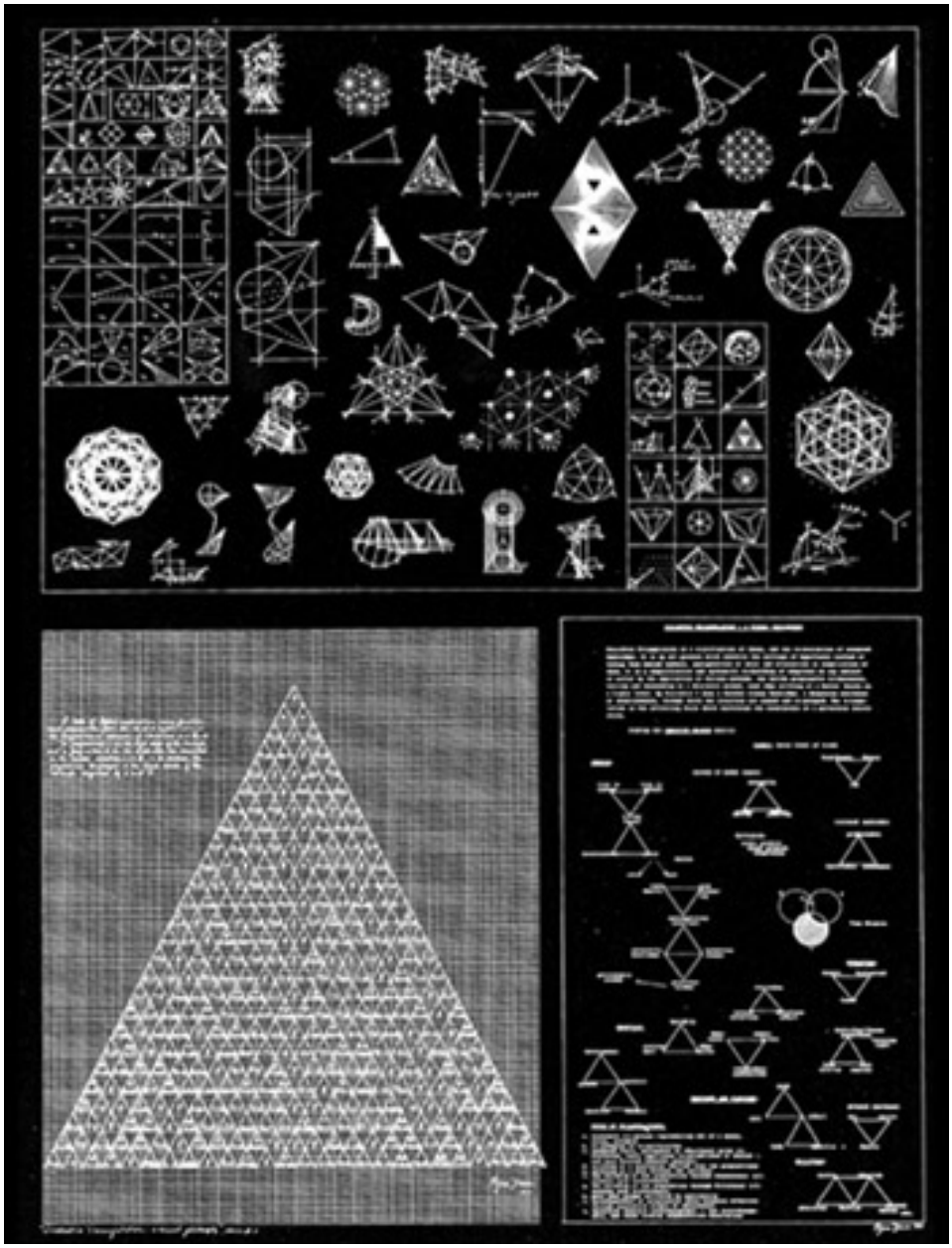


Figure 2.8. Agnes Denes, Matrix of Knowledge

Concepts can also be organized for aesthetic purposes. American artist Agnes Denes works with our metaphysical relationship to mathematics. In her work the Matrix of Knowledge, Figure 2.8, she explores a number of two and

three dimensional mathematical forms. While the presentation is semi-systematic, there are few labels and explicit relationships are unspecified. Instead the overall experience is one of mystery and wonder at these mathematical forms and a sense of awe with respect to the complexity of nature.

It is interesting to note that none of these forms are digitally produced or photographed. While there is a clear influence from science and mathematics, these drawings and their arrangement is not intended to be an objectifying one. Rather, the work forms a symbolic relationship with the viewer in deeper appreciation of a natural cosmology.

2.3 Computational Knowledge Organization

With the introduction of the computer it is natural that knowledge organization would move into the digital world. Due to the capacity and persistence of digital storage the computer allows knowledge organization to move forward more dramatically as it is no longer tied to the limitations of manually constructed diagrams and print reproduction. With the computer, it is much easier for multiple people to contribute to a system of knowledge. The computer is also capable of automatically generating diagrams, organizing concepts and providing screen-based interfaces to centrally located

resources. While systems of knowledge organization, such as the internet, will be explored further in chapter five (Systems) some specific notable examples will be presented here.

There are many ways in which computers enhance mental activities, including information management, numerical simulation, interactive design and communication. For the purposes of this thesis, Computational Knowledge Organization is defined as follows:

> Computational Knowledge Organization is the digital representation, storage and visualization of semantic concepts and their relationships according to type, class and specific properties for the purpose of clarifying and communicating these relationships to others.

If we append the clause "and for the economic benefit of a specific corporation or institution", then we have *knowledge management*. Corporations that provide important social services often have a need to manage large numbers of people, assets and processes. Assuming a flexible, semantically-based knowledge organization system were possible one must question the social impact such a system would have were it to be available only to certain institutions. The author holds that future information systems should be publicly available, relying on publicly funded free data sources with institutional profit derived from services rather than specific information and

technology. For this reason, the issues of profit, management and corporate development of knowledge systems will not be addressed in this thesis.

2.3.1 Early Database Systems

The first major success in using computers as knowledge organization tools came with the *database*. Prior to this, all interaction with the computer was through sequential instruction and punch cards. Two pioneers in this field were Charles Bachman and E.F. Codd. In 1969, the former developed the first *network model* for databases which would be become the foundation of the COBOL computer language [2-14]. The later developed the first relational model for databases, in 1970, which was used in the Informational Management System (IMS) developed by IBM [2-15]. Interestingly, the first application of this system was to inventory the extensive materials list for the Saturn V moon rocket.

Codd also distinguished between the *schema*, the class structure of a database, and its storage mechanism. This is crucial because the schema then becomes an abstract set of concepts that can be implemented independently of the particular hardware used.

Naturally, the first content to be migrated to the digital medium was that which was already becoming unwieldy in its physical form. In 1965, twenty years after Vannevar Bush's idea for a universal *memex*, J.C.R. Licklider wrote *Libraries of the Future*, outlining the research needed to build a working digital library [2-16]. The first successful Machine Readable Catalog (MARC) was implemented by the Library of Congress in the early 1970s. In a more recent book, William Arms outlines the many advances made in digital libraries over the years [2-17].

Throughout the 1970s and early 80s, interfaces to databases and library catalogs remained mostly textual. However, advances in *graphical user interfaces* were also proceeding in parallel to these developments.

A unique early example is the On-Line System from the Augmented Human Intellect project by Douglas Engelbart at Stanford in 1968. This was the first system to use the mouse and hypertext links to navigate information organized by relevance [2-18]. This project led directly to research done at Xerox PARC to develop the first desktop computer interfaces based on the windows, icon, menu, pointing device (WIMP) paradigm. The represented the first direct analogy between physical and digital organization [2-19].

Web browsers were the next major advance, connecting the desktop interface to the data of the World Wide Web. Pioneering systems for visual interfaces were also developed in the fields of information visualization, human-computer interaction and design. For historical context these will be explored first. Then we will examine the few (but growing number of) purely semantic systems whose goal is knowledge-driven rather than data-driven.

2.3.2. Scientific Visualization

Visualization, as a general term, is normal associated with the *cognitive* process of gaining insight and understanding. In this respect, it is independent of any specific data or machine representation [2-20]. Scientific visualization, as an instance of this, may be defined as the use of visual representations of data in a way that allows the human visual system to better understand the data as a whole.

The goal of scientific visualization is to leverage the cognitive abilities of the human visual system to see patterns in data where none were previously known. From the standpoint of communication, a visualization should convey a clear picture or representation of some scientific or natural phenomenon in order to promote understanding, discovery and research [2-21].

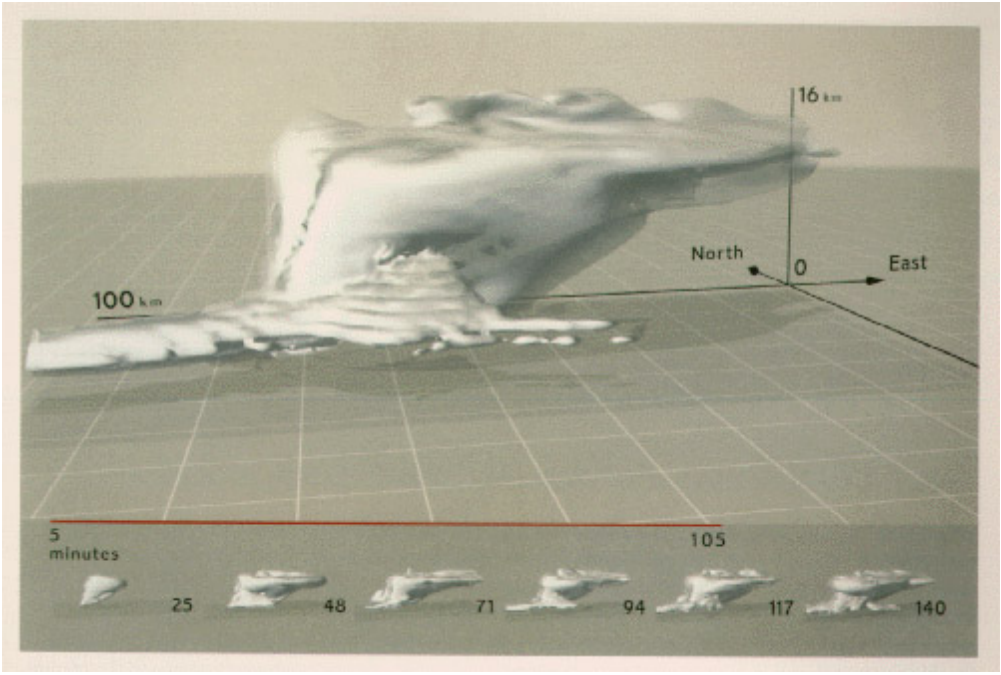


Figure 2.9. Visualization of a storm cloud, Tufte, 1983.

This image of a storm cloud by Tufte, found on the cover of his book *Visual Explanations*, is a clear example of quantitative visualization (Figure 2.9). We are immediately presented with the scientific experience of the storm cloud and its structure. The smaller icons show us how the structure changes over time. The goal is to elucidate the physical phenomenon of the cloud as present in the data.

Another example of scientific visualization comes from biochemistry. The Src protein is a complex molecule whose purpose is to regulate many aspects of cellular physiology [2-22]. Figure 2.10 shows a visualization of the Src protein from the Scripps Research Institute. Protein structures are usually derived

from X-Ray crystallography and theoretical models. However, X-Ray crystallography results in data that cannot be mapped directly to a spatial configuration. One must infer the structure from the diffraction of X-Rays [2-23]. In order to understand the result, it is helpful to visualize the final spatial structure. Thus information visualization is an important tool for examining results which have a physical interpretation but whose data exists in some other form.

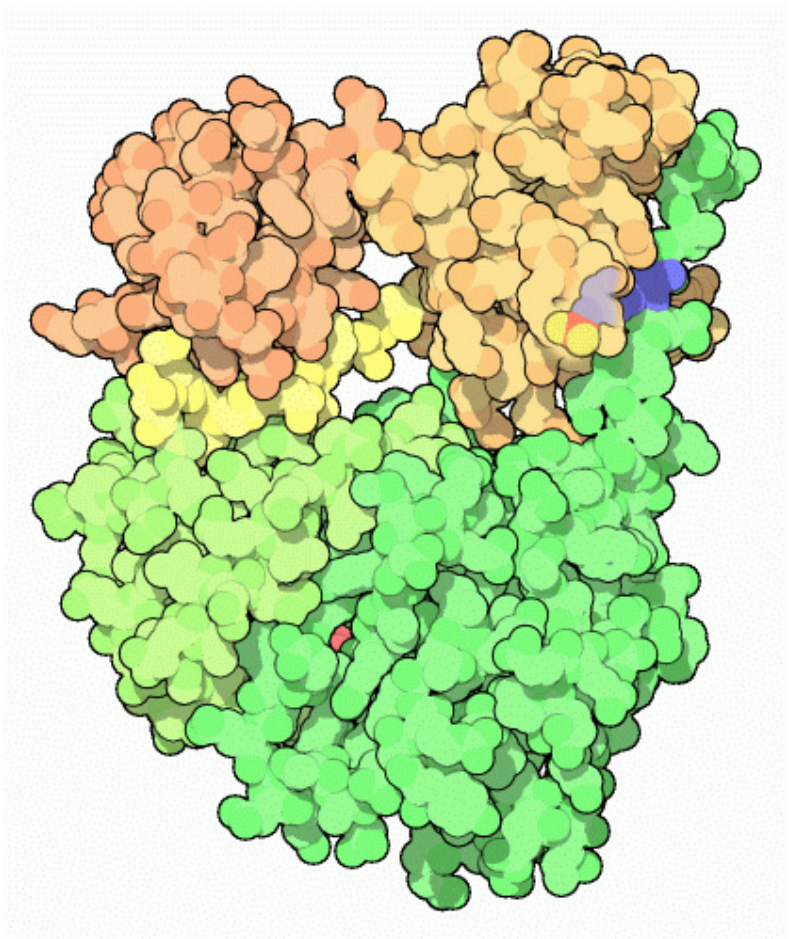


Figure 2.10. Src Protein, David Goodsell, Scripps Research Institute

Scientific visualization is a unique field in that it touches on many other disciplines. These include the natural sciences, visual design and human-computer interaction among others. While relatively new as a discipline, the ability to see scientific results visually has already had a significant impact on biology, chemistry, physics, economics and other fields.

2.3.3. Information Visualization

Information visualization, unlike scientific visualization, deals with the understanding of abstract relationships. Such relationships may exist in any conceptual domain and, since they may be arbitrary, information visualization focuses on techniques for presenting abstract structures like lists, trees and graphs. With these structures one may visualize any set of concepts.

Early work, for example Cone Trees [2-24] and Treemaps [2-25], allows for abstract navigation of trees. More recent approaches, such as Tamara Munzner's H3 in Figure 2.11, explores how large trees may be represented using hyperbolic geometry [2-26]. Here, space is distorted to place greater attention on the selected object at the center of the sphere. More distant concepts appear closer to the surface. This layout allows the user to more easily see and navigate large hierarchies.

Figure 2.11. H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space, Tamara Munzner (c) 1997 IEEE

The use of cognitive maps, as opposed to geographic maps, is another important aspect of information visualization [2-27]. An example in Figure 2.12. shows a map of chemical reaction pathways in a human cell. The base plane is a diagram for the various proteins in the cell. Vertical bars are used to indicate the levels of these proteins at a given moment in time. As time proceeds, the bars change height to indicate changes in protein levels as they might occur in a live cell. Conceptual maps thus allow us to see abstract relationships as well as change in time.

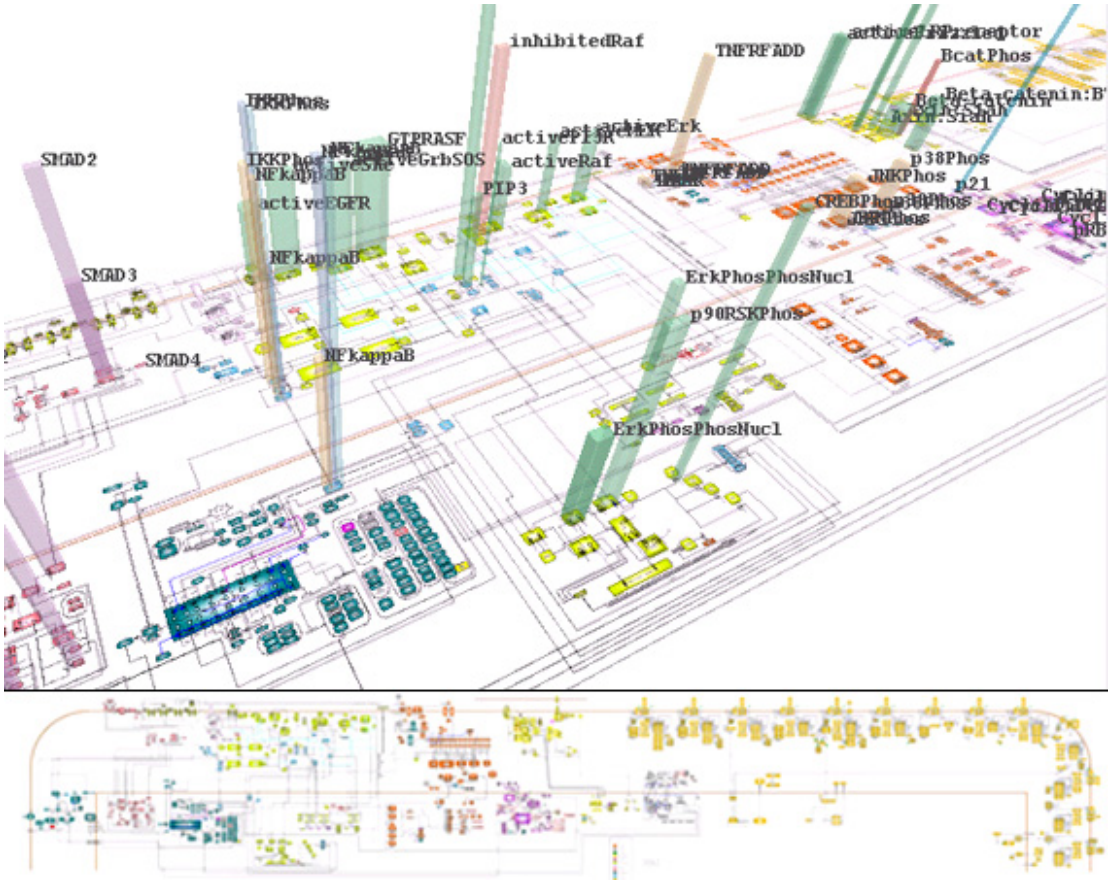


Figure 2.12. CellVis, visualization of protein levels in a human cell for a since moment in time. Visualization software developed by the author with Gene Network Sciences. (c) 2002 Gene Network Sciences, Inc.

Notice the difference between the Src protein in Figure 2.11 and the cell pathway map in Figure 2.13. In the first case, spatial positions map directly to a physical interpretation even though no direct photography was used. In the second, spatial position is used to arrange concepts but have no real physical meaning. This is the essential difference scientific and information visualization, and between geographic mapping and cognitive mapping.

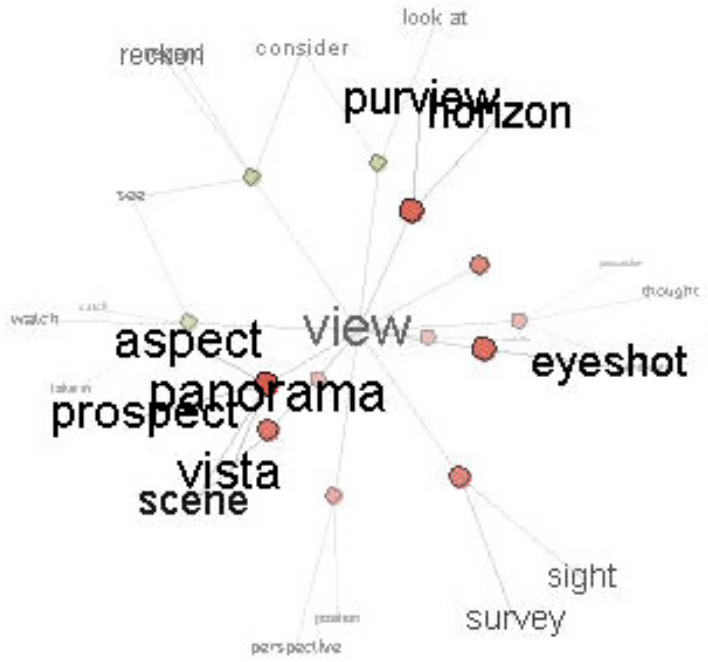


Figure 2.13. The Visual Thesarus, Thinkmap, 1998.
This image was generated by or is from the Visual Thesaurus (c) Thinkmap, Inc.
All rights reserved. More information available at: http://www.visualthesaurus.com.

Science is not the only area where information visualization may be used. A more practical tool for the individual user is the Visual Thesaurus developed by Thinkmap in 1998, Figure 2.13. Their software uses a dynamic layout that reconfigures itself, allowing the participant to navigate various word synonyms. As the navigation proceeds words expand and collapse to follow a stream of thought. Due to the document-centric nature of the web, information visualizations are not yet ubiquitous with the general public. However, with properly designed interfaces, this may change in the coming decades as visual navigation is a more natural way to navigate large-scale data.

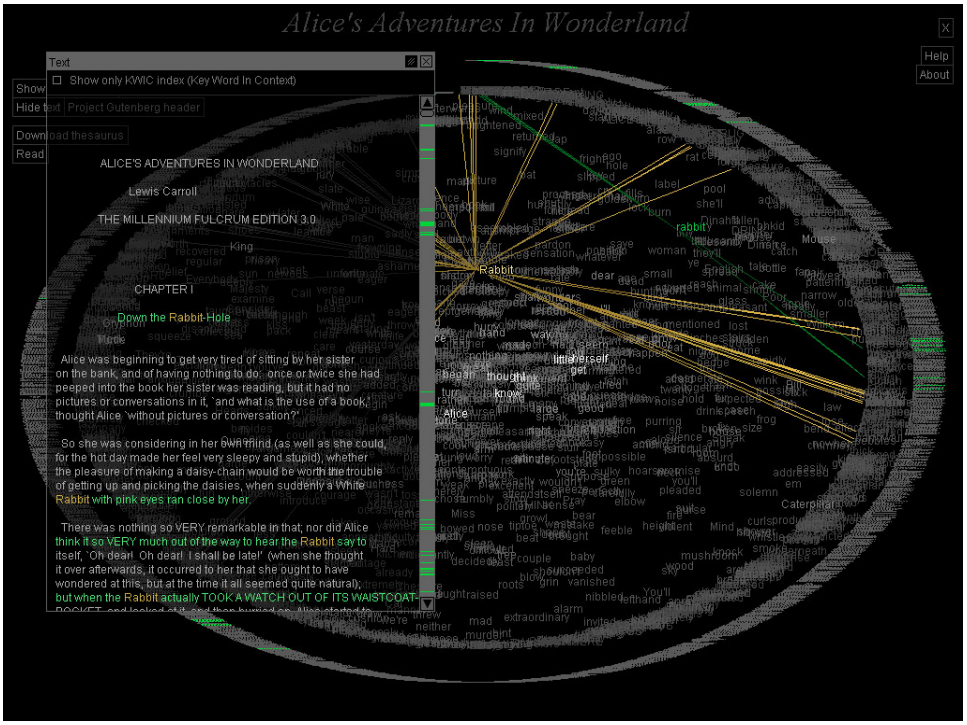


Figure 2.14. Alice's Adventures in Wonderland, shown in the tool TextArc, Bradford Paley (TextArc.org), 2002

2.3.4. Information Aesthetics

While scientific visualization deals with observable data, and information visualization with abstract concepts, developments were also taking place simultaneously in the visual arts through more social inquiries. Information aesthetics relies on similar tools as information visualization. However, the goal here is to to communicate a social or reflective message in addition to or instead of providing an understanding of the data [2-28].

The motivation for information aesthetics is not to solve a specific scientific problem but to convey a personal awareness or experience. There are no strict rules for the distinction between these disciplines but a straightforward explanation is that the former is driven toward scientific results, i.e. a novel experience leading to scientific realization, while the latter is driven by an aesthetic communication.

One example of information aesthetics is Bradford Paley's TextArc, shown in Figure 2.14. It is a visual representation of a document, in this case Alice In Wonderland, in which the entire text is composed as words arranged in a circle. The words are interactively connected in the order in which they are read on a set of arcs, commenting on the nature of text and writing as narrative [2-28]. Paley's system is also used as a tool by universities to allow viewers to navigate the original textual context in which related concepts are found.

Another example is Valence by Benjamin Fry, Figure 2.15. In this system, words are positioned three dimensionally with most frequent words appearing on the outside of concentric shells. Individual interactions between words are represented by dynamic curves connecting them. Here, Fry is interested in how aspects of biology help to provide an organic picture of dynamic data [2-29].

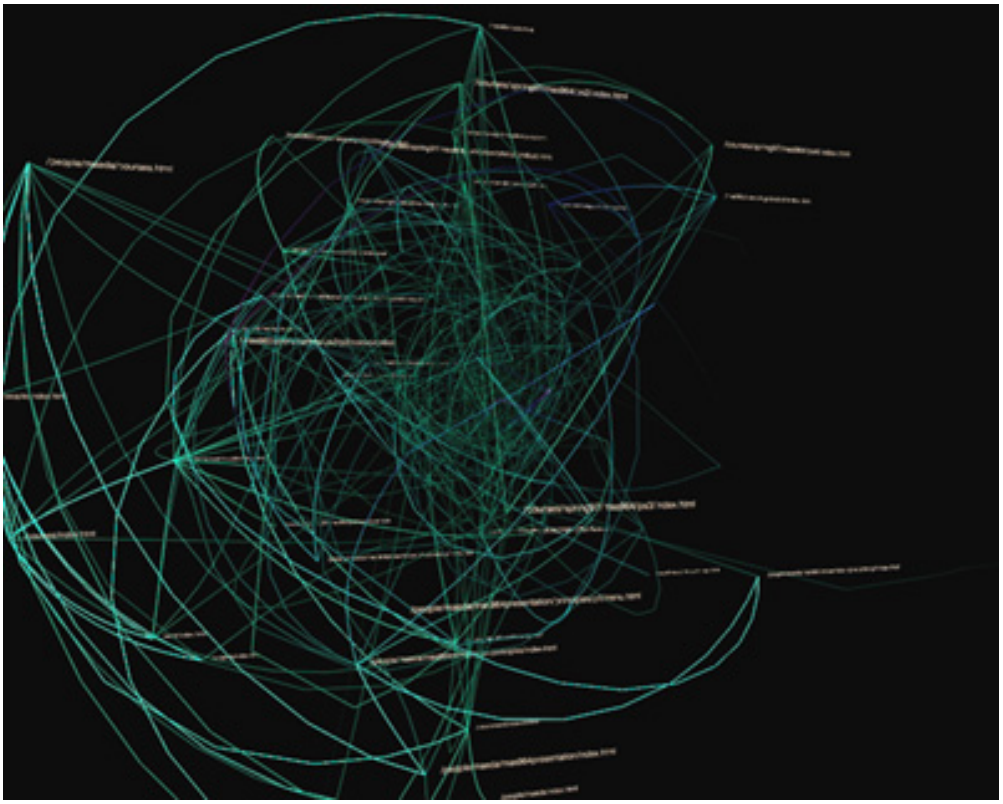


Figure 2.15. Valence, Benjamin Fry, 2000

The Seattle Library Visualization Project, by George Legrady, with technical production by the author is an ongoing visualization of current circulation of library materials at the Seattle Public Library. In Keyword Map, one of four visualizations, the most common keywords in titles checked out are arranged on a colored version of the Dewey Decimal Classification System (Figure 2.16). Position is determined by a weighting of the frequency a particular keyword appears in the various Dewey classes. In this system, the goal is not to come to a scientific conclusion but to provide a medium for social reflection on checkout patterns as patrons of the library move through the space.

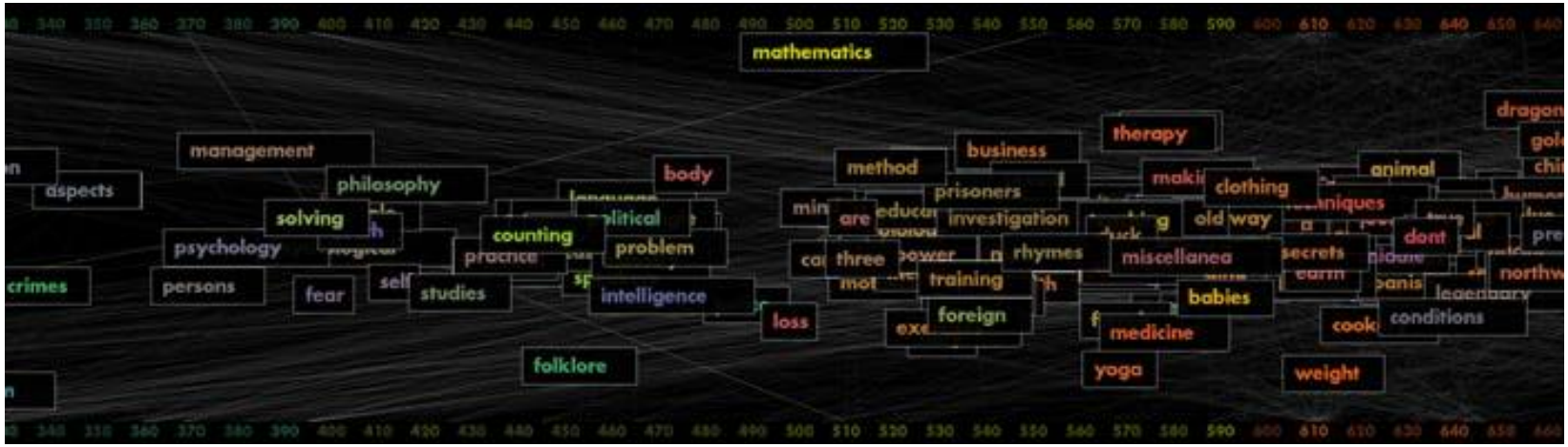


Figure 2.16. Seattle Library Visualization, George Legrady with Rama Hoetzlein, 2005

Work by Patrick Vuarnoz titled Studyscape, Figure 2.17, joins lecturers and students with their educational and research activities. With radiant circles and shades of blue the system takes on a highly designed quality. This is an important theme in information aesthetics when seen as a connection between futurist technology and reality. As a visual work, Studyscape gives a unique experience of space and distance while navigating ideas. Smooth animation contributes to the experience.

Figure 2.17. Studyscape, Patrick Vuarnoz, 2005

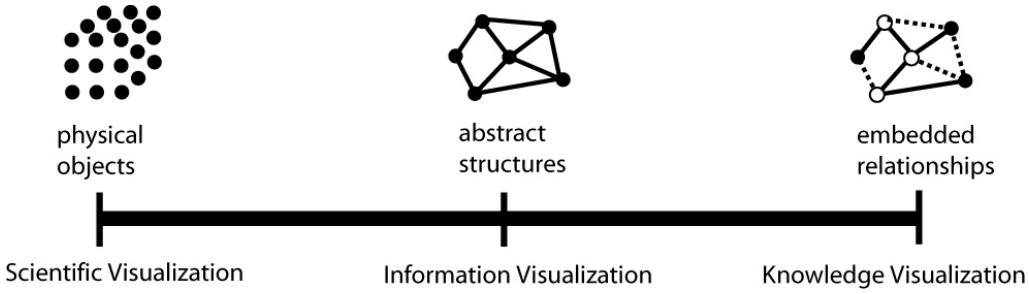


Figure 2.18. Continuum of scientific, information and knowledge visualization

2.3.5. Knowledge Visualization and Design

When we examine information visualization and design, we can observe certain patterns in approach. For example, when doing research in these areas the fundamental unit is typically a dataset with a specific relational structure. Scientific visualization deals with datasets that hold a direct correspondence to physical interpretation while information visualization and aesthetics both deal with abstract data but with differing communicative goals. In all cases, semantic relationships in the data frequently determine methodology.

Much contemporary research in information visualization is concerned with the structure of data and not necessarily relational complexity embedded in those structures. Knowledge visualization is defined here as the goal of understanding and visualizing rich semantic relationships in information

associated with language, i.e. to move beyond the use of particular relational structures. This does not eliminate the visual or structure component but suggests that new approaches are needed for semantically-rich data sets.

When dealing with large databases, one data element frequently looks very similar to another. However in semantically-rich data, the relationships are more important than the containing structure. It might be easy to assume that we could simply reduce these bits of data to an underlying network structure, but then we would be unable to explore the embedded relationships in a meaningful way. Simplifying a semantically-rich network as a graphs of nodes is similar to trying to understand human thought by looking at individual neurons.

To construct a more quantitative definition, I will define the *Data-semantic Ratio* for some set of information as follows:

> *Data-Semantic Ratio:* This is the ratio of the number of attributes of each individual element to the number of elements in the whole for some set of information.

This definition is largely inspired by a comment made by Noam Chomsky regarding language in which he states:

> "One's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like." [2-30]

We can take this to mean that individual words and sentences have a definite relationship to one another which, combined, give us an experience that is *knowledge*. However, statistics is necessary in order to resolve the many ambiguous meanings an utterance may have. We might consider these complimentary approaches. On some level each is true, statistics being a functional approach to disambiguation while the final interpretation of a sentence is a specific thought [2-31].

Information visualization deals with conceptual structures based on the structure of the data in its machine representation. As a bottom-up approach, this may take the form of trees, graphs and networks. Yet *thoughts* have much more semantic content than this. The continuum between these two views is captured by the Data-Semantic Ratio (DSR) for measurable systems. For example, in the storm cloud of Tufte the information about each point is small, perhaps consisting only of volume, pressure, temperature and density - say 10 attributes. Yet the total number of data points is very large, perhaps one million. Thus the ratio in this example is 1 / 100000. The model has a low data-semantic ratio.

There are few examples of systems that deal with information with a large Data-Semantic Ratio. One example, however, is a digital encyclopedia such as Wikipedia. The average number of words per article is roughly 400 while the total number of English articles is 883,000 as of 2006. Thus the ratio is 1 / 2200. [1] The primary drawback with digital encyclopediae is that their interfaces still follow classical article-based models rather than relying on visual design.

It will be found in chapter seven that most current systems for information visualization deal with only one workable element at a time (i.e. the person, the molecule or the article), then apply visualizations to the attributes of that element. Knowledge visualization is the goal of applying design principles to semantic data without breaking the complexity found at different *scales* so that visual structures are not tied to the internal representations of the data.

Interestingly, we may find that what we call "physical things" will have a very low DSR (simulated water, clouds, fire) because they have many data points with few variables at each point, while what we call more "mental things" will have a very high DSR (words, concepts, abstractions) since they have few abstractions on the same level but many facets within each abstraction. In

[1] A better measure would take into account the scale-free nature of these systems. Wikipedia has very many articles but most of these have only a few words. The most read articles have a large number of words.

mathematical terms we could say the Data-Semantic Ratio is a linguistic fractal dimension[2]. Are there more things in the world? Or more qualities in one of these things?

Obviously, for a theory of knowledge visualization be successful there must be a shift from data sets (small DSR) to semantic systems (large DSR). While information visualization relies on a visual *reduction*, knowledge visualization is about retaining that complexity by placing the viewer "in the data" through navigation. Design methodology shifts toward understanding the *meaning* present in the information, not just its structure.

Scientific visualization is an abstraction from raw data to spatial visualization. Information visualization abstracts this to visualize not only spatial data but arbitrary concepts based on data structures. Knowledge visualization abstracts away the *structure* of data to focus on navigation and meaning. Structures and visual representations are not eliminated, but rather than focus on the structures themselves, design effort is placed on the natural flexibility of thought, navigation through multiple structures, and the dynamic reconfiguration of structure.

[2] Where fractal dimension is classically defined as the ratio of the number of self-similar pieces to the magnification factor, ie. the scale of parts relative to the whole