Chapter 1

Introduction

The organization of knowledge is a uniquely interdisciplinary activity in that its problems require the input and expertise of many disciplines. The issues involved are rich and complex, including the study of perception, language, cognition, metaphysics, philosophy, semiotics and vision. When machines are introduced to solve these problems the scope expands to include database systems, information visualization, software design and distributed computing. To facilitate these discussions the first section presents an overview of the structure of this thesis from both a conceptual and technical perspective. Quanta, an integrated software system for generic knowledge organization is introduced.

1.1. Overview of the Thesis

The goal of this thesis is to understand and address issues in knowledge organization from an interdisciplinary perspective - to integrate the historical context and current understanding of knowledge organization in various domains in order to arrive at a unified, balanced solution to the social classification and arrangement of knowledge. This is to be achieved through

an analysis of a number of fields including computer science, philosophy, linguistics and information visualization.

This work originally began as a series of thought experiments in 1998 in an attempt to answer the question: What are the limits of a machine in the simulation of complex objects and ideas? Vines growing on walls, building foundations buried in sand, cities rising from deserts, water flowing in ravines, and the complex structures of the urban landscape may all be imagined to provoke the limits of this question.

Interest in machine representation led to the conclusion that this was not only a problem in space, geometry and physics but also in semantics. How does one express the partial replacement of the bricks of an older wall with a newer construction - simultaneously entwined by a nearby tree? Such descriptions are difficult even in natural language. Unlike the rigid bounds of words and disciplines, the types of complex relationships found in an ancient wall through arbitrary subtraction and addition in time are not the exception but the norm in most fields of study. Biology is intertwined with ecology and chemistry, and chemistry with physics. Engineering requires an understanding of manufacturing and economics but is also a study in design, which has its own origins in art. The motivation of this thesis is to be able to see and navigate these relationships freely. To help dissolve the artificial

distinctions between disciplines that have been created over time through a fluid reorganization of knowledge.

A conceptual overview of this thesis is shown in Figure 1.1. This map shows lines of thought with the relevant chapter indicated next to each general area of discussion. The connections reflect the general development of the project from its conceptual origins. Dates are not shown since many ideas were revisited over and over as other ideas informed them. Specific novel contributions are indicated in boxes with details found in their respective chapters.

The goal of this thesis is twofold. The first is to understand contemporary issues in knowledge organization and to resolve key problems through interdisciplinary analysis and specific solutions in each domain. The second is to present Quanta as a novel system for the study and exploration of knowledge. While many other systems already exist, both traditional and modern, this system is designed from the beginning on the principle of an interdisciplinary integration of techniques from multiple fields.

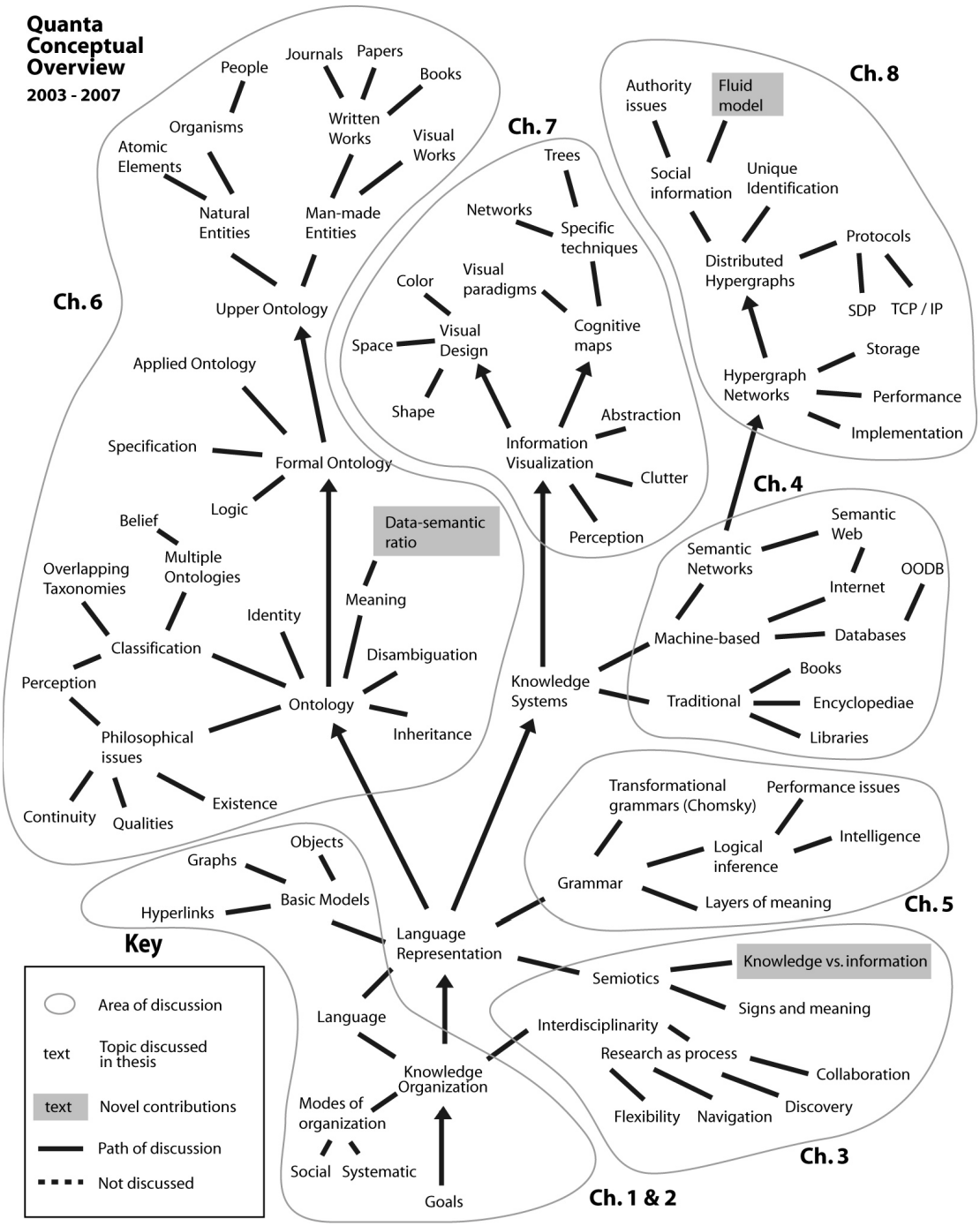

Figure 1.1. Quanta conceptual overview.

Quanta is presented here as a prototype for the implementation of theoretical results made in this thesis. Novel solutions in database design, information visualization and ontology design, discussed in detail in the following chapters, are combined in Quanta to present a system developed on the principle of a balanced synthesis of these tasks. Six visualizations allow a user to dynamically navigate and explore complex semantic data through a variety of modalities. The test data set was constructed incrementally from different sources in a number of disciplines. With the development of a novel semantic database, interaction in Quanta allows for more precise navigation of concepts and ideas than is available with an article-centric resource such as the current internet. Quanta is developed as a novel research tool that allows users in different fields to explore concepts outside their expertise but in sufficient depth and detail to be meaningful. Implementation details, limitations and future directions of the system are discussed in chapter nine.

The development of Quanta and relevant chapters to its design are summarized in Figure 1.2. Programming of Quanta began in 2003 with a custom database written in C++. Some design decisions were backed by theoretical arguments, others by existing approaches or by time constraints. Later on many ideas were eliminated based on balancing goals across the whole system. Grey lines show ideas that were abandoned, solid lines paths that were pursued, and dotted lines potential areas for future growth. The

larger modules of the system are drawn as regions. Dotted regions are considered but not yet implemented. Considerable effort in implementation was spent on items in bold while items in boxes represent novel contributions described in their respective chapters.

In the next section, novel contributions of the thesis are presented for reference and to provide a more detailed overview of this work.

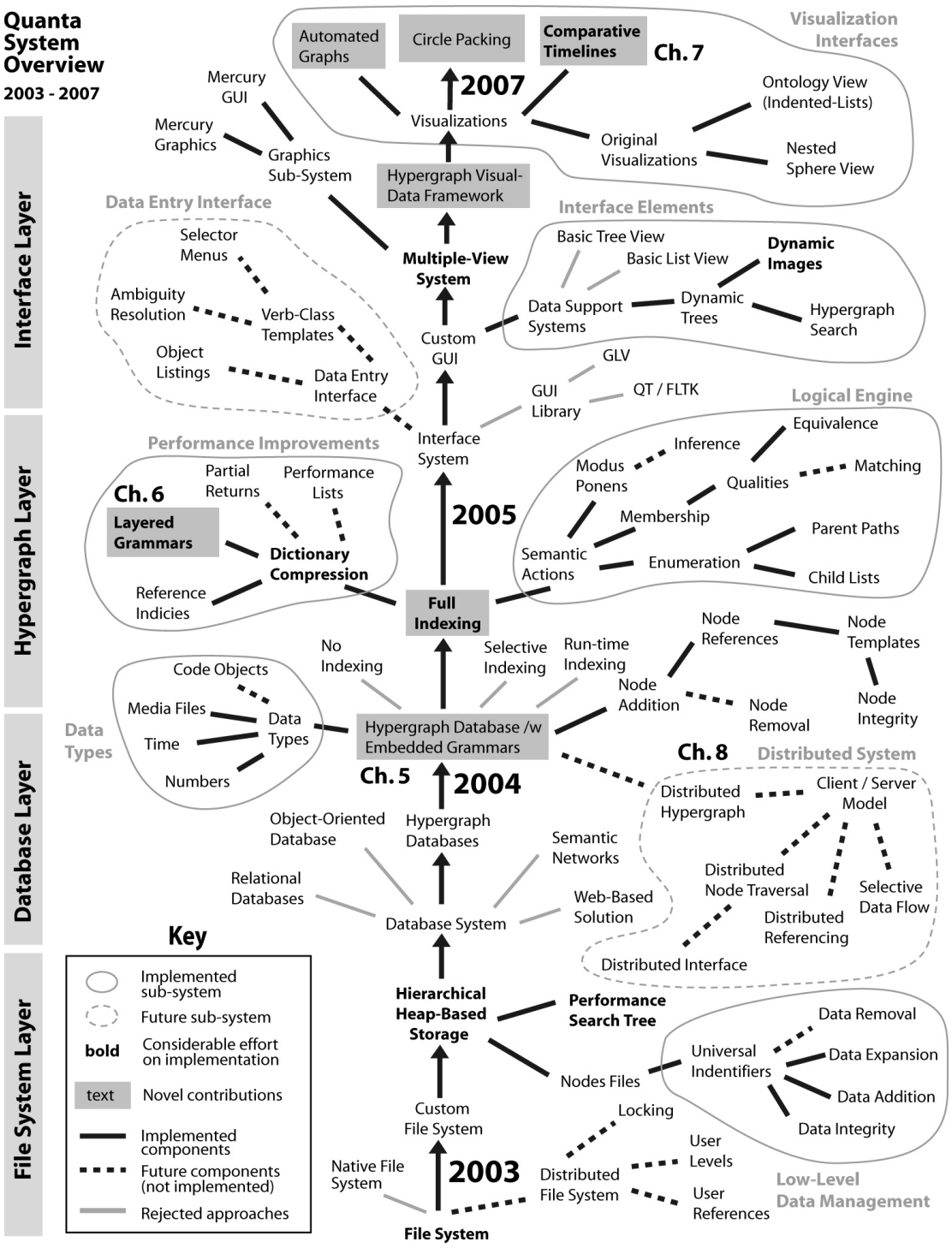

Figure 1.2. Quanta system overview.

1.2. Novel Contributions

Discussions in this thesis will cover a wide range of topics. While a broad shift in perspective is advocated across disciplines, it may also be helpful to distinguish specific novel contributions made in each area. These contributions were developed at various times during the completion of this work from 2002 to 2007.

● The Data-Semantic Ratio (DSR) is introduced as a novel technique for measuring the meaning content of a knowledge system (Ch. 2).

● Knowledge Visualization is distinguish from Information Visualization through arguments regarding the flexibility of human cognition (Ch 2).

● New definitions are introduced for Data, Information and Knowledge by way of comparison to previous definitions and with the use of a thought experiment (Ch 3).

● A unique relationship is demonstrated between the semiotic triangle (referent, signifier, signified) and the data-information-knowledge triangle through a comparison of semiology to information science. (Ch. 3)

- The Quanta software is introduced as a prototype for the novel design of a knowledge system in which the aspects of storage, representation, language, ontology and visualization are treated equally during the design phase. (Ch. 4)

- A novel hypergraph database architecture with a layered grammar is introduced to store and represent knowledge. This design combines graph-oriented databases, hypergraphs, and semantic networks into a unique system that functions both as a database and as a grammatic structure. Full indexing allows for efficient induction and querying while the system remains linguistically flexible. (Ch. 5)

- Investigations in ontology design reveal that while multiple hierarchies over a set of qualities of an object allow for faceted classification, multiple hierarchies on a single quality are necessary to express belief (Ch 6).

- The *comparative zoomable timeline* is introduced as a novel visualization of events in time. This view-dependent system allows events in multiple disciplines to be compared while also providing continuous zooming in both time and level of detail. (Ch. 7)

● Circle packing for the visualization of large trees, while not novel itself [7-14], is extended to provide continuous navigation with dynamic loading of data from a supporting hypergraph database (Ch 7)

● A novel design is presented, but not implemented, for a distributed semantic network using a new low-level protocol, the Semantic Data Protocol. This protocol, while experimental, should allow for greater efficiency in storage and communication of concepts on a semantic network than could be provided by strategies using metadata. (Ch 8)

The above contributions are presented to support a general shift in perspective away from discipline-oriented investigations of knowledge toward an integrated, systemic approach. With the exception of the last item, all technical contributions above are incorporated into the Quanta prototype system. Conceptual contributions, such as the data-semantic ratio, are made by way of example, historical context, and general discussion.

The primary contribution of this thesis is to present a strategy for an integrated approach to knowledge organization. The goal is to offer a shift in perspective that unifies our approach to knowledge organization through a synthesis of concepts. This refocusing of effort toward integration always competes for time with the production of narrow, specific results in one

discipline. It is necessary to be selective on the amount of detail in certain topics in order to achieve this synthesis.

Furthermore, the broad impact of integrative research is not necessarily measurable on any standard scale. As with other fields in the humanities, long term strategies to synthesize fields are perhaps most valuable but least susceptible to quantifiable short term outcomes. Therefore, to engage in interdisciplinary studies of knowledge requires a change in how we do research, but also a change in how evaluate, publish and reward that research. The concept of measurable value is persistently ellusive as a metric for interdisciplinary methods. Any field of study is a story of connections extending back in time and ultimately connected to all other disciplines.

To engage in interdisciplinary research is to focus on the bigger picture in addition to the details. The humanities often struggle with the details of technology while engineering disciplines struggle with context and larger social meaning. The concept of balance in methodology is not found in breadth or in depth, but in both simultaneously. This is achieved through careful selection of topics. This thesis will present detailed results in specific fields set in the broader context of a synthesis of knowledge organization.

1.3. Types of Knowledge

Epistemology, from the Greek word *episteme* (to know), is the branch of philosophy that deals specifically with knowledge and how it is acquired. In Plato's ancient Greek work the Republic, a divided line is used to illustrate the difference between illusion, belief, reason and intelligence (Figure 1.3). With this system Plato formulated one of the first distinctions of knowledge into categories based on degrees of *truth* [1-1].

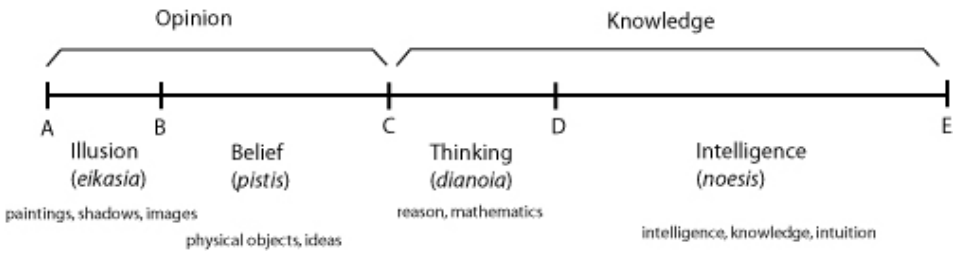

Figure 1.3. Plato's Divided Line used to represent the forms of knowledge.

This distinction between experience, belief, and pure knowledge is the foundation of *rationalism*. This may be compared to *empiricism*, originated by another Greek philosopher Aristotle, which is founded on the idea that all things may be explained through experience or perception [1-2]. These two systems of truth are the foundation of contemporary dialogues in Western philosophy.

Discussion on how we identify true knowledge and its origins continues to the present day. More recently, newer schools of thought include *materialism*, *pragmatism*, and *analytic philosophy* which present the views, respectively, that knowledge always has a material substrate [1-3], that our approach to knowledge is limited by language and human experience [1-4], and that the only valid knowledge is that which is objectively reasoned [1-5].

This thesis, however, is not concerned with the evaluation of knowledge in relationship to truth but with the *organization of knowledge.* In this respect the organizer of knowledge is not so concerned with correctness of a particular idea, but with placing it in a meaningful location relative to others. The librarian does not weigh the relative merits of one book over another but places both on the shelf of knowledge. While it remains essential for humanity to think, reason and evaluate the truthfulness of a concept, the topic of this thesis deals with how we classify and organize rather than how we evaluate ideas. Thus the first task will be to formulate a clear definition of knowledge organization.

Organization, as a principle, can be found in Plato's descriptions of Socrates in identifying the need to "perceiving and bringing together in one idea [form] the scattered particulars, that one may make clear by definition the particular

thing which he wishes to explain." [1-6] A more modern definition of knowledge organization can be found in Berwick Sayers who describes it as:

> "not only the general grouping of things for location or identification purposes; it is also their arrangement in some sort of logical order so that the relationship of the things may be ascertained." [1-7]

While this gives a broad definition the pressing question is the best *logical order* for revealing these relationships. The most obvious is subject classification in which the boundaries of various disciplines are defined according to their area of study. But is this the best way?

The breadth and depth of knowledge is vast. Some knowledge may be true and some false (within a school of thought). Some knowledge may be fantastical, such as descriptions of dragons, while other knowledge may be empirical, such as a description of the moon. Some knowledge may be more universal, such as the laws of gravity, while some may be temporary, such as current trends in popular music. Finally knowledge can be very specific, such as the names of the thousands of proteins that make up a human cell or the gene sequence of the human DNA.

Knowledge found in various disciplines is not simply a difference in content but may be fundamentally different in nature as well. For example Tacit Knowledge, first described by Polyani, expresses the idea that we carry the knowledge of certain rituals and tasks inside us in ways that cannot be

expressed in any language [1-8]. No matter how many times I describe the process of riding a bicycle you will not learn how to actually ride one from this description. Thus fields of study often follow a *methodology,* not simply a variation on content but a different way of *knowing and pursuing* knowledge.

One theme of this thesis will be to establish what principles of *order* are best for the classification of general knowledge. While general schemes will be examined we should keep in mind that that perhaps no universal ordering can be found. In chapter six (Ontology and Classification), the problems with universal classifications and their alternatives are explored further.

1.4. Libraries

The task of collecting, classifying and organizing knowledge was historically the role of the philosopher but has a more conventional profession in the methods of the librarian. The original task of the librarian was simply to record the holding of a library in a *catalogue*. While not the first one of the great ancient libraries was the Library of Alexandria, which reached its peak in the 3rd c. BC. With over 500,000 scrolls, it was maintained by Demetrius Phalereus under the direction of Aristotle. In addition to sorting scrolls according to their contents as Aristotle did, the first appointed Director of the library, Zenodotus, organized the titles in alphabetical order. While only the

first letter was alphabetized, it was not until the 2nd c. AD that newer methods for classification would appear [1-9].

The most influential modern advancement in library classification may be the Dewey Decimal System developed by Melvin Dewey in 1876 based partly on notes by William Blake [1-10]. This system, like the ancient ones, divides knowledge according to subject area. The unique aspect of the decimal system is its ability to be infinitely divisible by using decimal numbering. The top-level classes are organized into groups of ten, which are further divided into one hundred categories each. Beyond that, decimal numbers are used to allow for an infinite number of sub-categories.

Yet the Dewey Decimal System is not the only approach to classification. There are several others including the Library of Congress Classification and Universal Decimal Classification, which use different numbering schemes [1-11]. One interesting alternative is Colon Classification developed by S. R. Ranganathan and used primarily in India [1-12]. This is called a *faceted* system because the call number is constructed to classify the material by certain principles.

These principles are identified by particular punctuation as follows:

, personality

; matter or property

: energy

. space

' time

Thus, we can interpret the call number in the following example:

L,45;421:6;253:f.44'N5

Medicine,Lungs;Tuberculosis:Treatment; X-ray:Research . India ' 1950

The interesting aspect of colon classification is that the call number itself appears to locate the concept relative to others: "Lungs" is the type of "medicine" we are interested in and "treatment" is what we wish to know about "tuberculosis".

Are library classification systems the best way to organize knowledge? While they each have their benefits there are also certain drawbacks. One problem is how to classify newer interdisciplinary materials. Consider synthesized bacteria as an example. Is this to be classified as a biological organism or an engineered object? Many physical library classification systems demand classification in just one category - so the item can be placed on the shelf.

Another problem with library classifications is their fixed, hierarchical nature. Advances in many areas occur so rapidly that it is impossible for classification

systems to keep up with new research. New additions to the field of mathematics include *levels sets*, *parzen density estimates*, *fractal geometry*, *set theory* and *cryptography*. Yet these cannot be found in newer classifications partly because they are so new, partly because only mathematicians can properly define their categories, and partly because they cannot be placed into a rigid hierarchical system (they may depend on each other).

Call numbers, such as those above, have their benefits and drawbacks as well. For physical libraries, they are ideal when some physical material such as a book must be located at a specific place in the library. However all of this may be made somewhat irrelevant by the revolutions of keyword searching and the Internet in which there is no physical location [1-13].

1.5. The Internet

If the first revolution in knowledge organization was systematic classification the second is digital libraries and the Internet. With the internet conceptual resources no longer need a physical location. In addition, search engines such as Google allow us to easily find materials without the need for call numbers. We simply enter several keywords and matches appear immediately. In terms of common knowledge, the Internet is nothing less than

a revolution. Despite this, we might ask if the way we do research changed that dramatically?

Vannevar Bush, in a paper called "As We May Think" in 1945, described a universal *memex*, a device which can "store all books, records and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility." [1-14]. Written even before the concept of computer networks, his memex foreshadowed the internet. Yet it is also more than that:

> "The owner of the memex, let us say, is interested in the origin and properties of the bow and arrow. Specifically he is studying why the short Turkish bow was apparently superior to the English long bow in the skirmished of the Crusades. He has dozens of possibly pertinent books and articles in his memex. First he runs through an encyclopedia, finds an interesting but sketchy article and leaves it projected. Next, in a history, he finds another pertinent item, and ties the two together. Thus he goes, building a trail of many items. Occasionally he inserts a comment of his own, either linking it into the main trail or joining it by a side trail to a particular item." [1-14]

The most sophisticated tool in wide use for the internet is the search engine. While it allows for complex, efficient retrieval it does not yet allow for the kinds of complex interactions, such as comparing or "joining" thoughts, as described by the *memex.* This can be traced to the fact that the fundamental unit of the world wide web, the global media content of the internet, is that of the website or article. Individual pieces of knowledge themselves, the sentences we read, are contained in these sites in natural langauge yet no

machine is capable of autonomously navigating the thoughts themselves as described by the memex.

The problem of *knowledge representation* is how to store concepts in a meaningful way so that their relationships can be made explicit. At present, connections on the Internet are based on the hyperlink which must be explicitly identified by the web author. Tim Berners-Lee, the inventor of HTML, has suggested that additional semantic information added to an authored web page, called *metadata*, would allow for a different kind of representation [1-15]. This and other ways of representing knowledge will be examined more carefully in chapter five (Language and Representation).

1.6. Integration

Let us look at another example of a knowledge driven research problem. Perhaps we would like to identify the evolution of birds and study corresponding relationships in the evolution of airplane designs. While this clearly falls under engineering, there may be some real biological aspects - genes that control aspects of flight or navigation - that can only be found in the study of birds of flight as biological animals.

Such studies may already exist, but do we search for them in biology or in aircraft engineering? If we are familiar with only one field the discussions in the others may be lost to us due to language. We may not be able to identify the impact the other field has on our own since we do not know where to look. Ideally, the system itself would be able to suggest when there is a corresponding relationship with another field. The problems of interdisciplinary integration of knowledge will be examined in chapter three (Integrative Strategies).

The Internet has certainly improved our ability to find unknown information quickly. Yet it is not difficult to identify specific tasks which we would like to perform as researches that are not yet possible. These include:

- Requesting a side-by-side comparison of advances made in abstract art and biology over the last fifty years.
- Locating a journal article, then asking for a visual tree of references that extend back to original authors on the subject as well as the current top five researchers in that field.
- Requesting a list of all mathematical algorithms first organized by date of discovery, then rearranged by most common application area, and finally organized by the class of problems they solve.
- Requesting a display of all images that contain a depiction of bees and the corresponding works and disciplines these images come from.
- Comparing the lives of two scientists in terms of the age of earliest publications and specific publications each year.

These are only representative examples. There are many other kinds of meaningful queries we might like to engage in. There is no theoretical reason why these kinds of questions cannot be answered by a computer. Many alternative systems already exist for doing simpler versions of the above queries. These will be examined more carefully in chapter four (Systems).

Some of the query examples above also lend themselves more naturally to visual interaction, such as performing visual comparisons of artworks by different artists and periods. For a medium that is capable of dynamic visual navigation, the internet is still largely a textual environment for research. Unique advances are also being made in this area. Examples of visual interactions with knowledge systems, and the issues they raise, will be explored in more detail in chapter seven (Knowledge Visualization).

Knowledge organization is a non-trivial task made more difficult by the specific, complex relationships in new types of knowledge found in modern disciplines. The problems to be solved draw on the fields of computer science, linguistics, philosophy, information visualization and natural science. Each of these must be addressed together to develop complete solutions. The problem of knowledge organization can therefore be viewed as a uniquely interdisciplinary problem whose solution is only possible through an integration of ideas and methods.