The Large Data Construct: A New Frontier in Database Design

Stephen E. Arnold

Information Access Co.
Foster City, CA 94404

Rapid advances in microcomputer processing power have accelerated the development of multi-object databases. These new information constructs require different record layouts, demand the inclusion on non-searchable strings, and place different demands upon the query logic in the software and in the mind of the searcher. Traditional databases are typically collections of fairly uniform records. These are usually abstracts, indices, and full text in a range of combinations. Such traditional databases are collections or small data constructs. The new databases combine text, images, recorded or synthesized voice, and other objects. Databases that contain multiple objects represent the data. Such representations are large data constructs. This new terminology reminds the database builder and the consumer of information that fundamentally different approaches to large data constructs are the only way to explore these radically different databases.

Berkeley Sunday: Hit Fat Apple's, check out the New York Times and Sun Jose Mercury News, ond flip through the most recent **Whole *Earth Review.*** I am sitting in the sun and have **three–yes,** three-current articles about hybrid databases (aka virtual realities or cyberspace). Coast-to-coast information constructs are real and hot.

Of course, writers for the dailies sniff out trends before the rest of us have a clue. That's why the front page of the New York *Times* blares in 36-point type: "Virtual Reality Takes Its Place in the Real World" (Hall, 1990). This article bounces between the gee-whiz and the don't-get-your-hopes-up-yet-buddy putdown of safe journalism. Author Trish Hall trots out Jaron Lanier, president of VPL Research, Inc. (Redwood City., CA), and godfather to the Mattel $90 Power Glove. Jaron's dreadlocks are a photo opportunity. Then Ms. Hall quotes Myron Krueger, a computer scientist who is "something of a maverick these days." He says: *"We* are all creatures of artificial experience....' The new technology 'allows that symbolic world to become concrete."' Expectedly the photographs accompanying the article feature Messrs. Lanier and Krueger wearing data gloves and holding their hands in quasi-kung fu gestures. Real reality is weirder than virtual reality in Times-land.

In an article appearing in the *Mercury* News, James Lalonde (1990) of the *Seattle Times* writes 'World is Virtually a Reality: Cyberspace Lab Explores Frontiers." Lalonde describes the University of Washington's Human Interface Technology Laboratory or HITL. Digital Equipment Corp. dropped about $1.4 million to pay for research and fund the "Virtual Worlds Consortium." The idea is to "explore radically new methods of human interaction with computers and massive amounts of data stored in them." One of HITL's spark plugs, Mr. Furness, is one of the engineers associated with the development of the heads-up stores display in a modern fighter aircraft. At HITL, Mr. Furness wants to use a laser to scan graphic images directly on the retinas of computer users. Mr. Lalonde quotes Andrew Zarillo of Autodesk (Sausalito, CA): "You've got to be careful about predictions." The last paragraph of the story reminds me that Canadian novelist William Gibson invented cyberspace in 1984 in the sci-fi novel *Neuromancer.* Let's see, that was six years ago, and stories about cyberspace are just now appearing. This is news?

In Sausalito's own *Whole Earth Review,* Howard Rheingold (1990), laboring on a book about the worlds created by software, does a thorough job of surveying the major players in this segment of the information industry. For aficionados of The Well online service, Mr. Rheingold manages the virtual reality forum.

Mr. Rheingold's first stop is the University of North Carolina at Chapel Hill. UNC has one of the older virtual reality research projects in the US. Mr. Rheingold names the pioneers in this field: J.C.R. Licklider (formerly of ARPA and SRI, and now at Stanford University) and John Walker (Autodesk). He describes Margaret Minsky's virtual sandpaper; a collaboration between MIT's Media Lab and UNC; two recent defectors from Autodesk and their homebrew virtual reality machine; the University of Washington lab; Jaron Lanier; Dave Johnson of TiNi Co. (Emeryville, CA); and wraps up with an extract from a position paper crafted by Bob Jacobson at UW's HITL. An extract from this article is my beginning:

> **A "virtual world" is a unique, intangible but highly designed information environment generated by a computer and transmitted by "virtual-interface" technology to a user who "enters" the virtual world via appropriate sensory mechanism. The virtual-world environment can be as complex as a three-dimensional "sense surround" comprising seamless visual, aural and tactile cues; or as simple as a computer conferencing system. Virtual worlds are designed to increase the bandwidth of communication between the computer and the human being, to facilitate their interaction, and ultimately to improve the human being's understanding and performance. The subject of this news group will be virtual worlds in all their aspects: the theory of virtuality, the technology that is being developed and employed to create virtual-world environments, the people and places working on virtual worlds, and the philosophical questions and social consequences attendant upon the emergence of this new medium of communication. (Rheingold, 1990, p. 87).**

Well, my approach is much less exciting than the examples from two newspapers and one magazine. The basic tool of the virtual reality is the microcomputer. Tomorrow's machine will be the United Airlines to these new data worlds. For information professionals, the information management issues associated with cyberspace are going to be thorny. The purpose of this essay is to hold up several ideas before you and I jack into cyberspace and do online research in a wonderful and wild way.

IA. A COLLECTION

Let's startwith a new way to think about traditional online databases. A database is a collection, and a collection is a group of objects or an amount of material accumulated in one location. Figure **1** (p. 188) shows one way to visualize a collection of similar objects. I can have a collection of hats or a collection of software. A collection can also be complete. But *completeness*-having every issue of *PC* Week, for instance-does not present the items in their original context. My stack of back issues doesn't show a reader scanning product comparisons or tearing an article out of the most recent issue.

The databases that my colleagues build at work or those cranked out by massive database production entities like the American Psychological Association or Engineering Information have more in common with a glass case filled with Lepidoptera than with the new electronic information constructs. Traditional database producers share the fanaticism of the late Malcolm Forbes for preserving items for posterity. Mr. Forbes collected toy soldiers, boats, and Fabergé eggs. We selectively collect index citations, article summaries, and full text newspaper articles.

Collections are classification schemes with real objects illustrating the conceptual framework of the collector. When we encounter a collection, we see specific examples of a particular classification scheme. In everyday speech, the word *collection* works well. No one is confused if I ask, "Do you want to look at my collection of audio CDs?"

I enter more dangerous precincts if I ask, "Do you want to see my collection of online abstracts from Compendex Plus?" Here my collection is an extract from a larger collection. In fact, Compendex Plus, the source of the collection, can never be complete. New material is added to the database everyday. When I access the database, I look for new records added to the Compendex Plus collection. Obviously my collection has different contents depending upon when I look at the online file.

If all this seems unnecessarily muddy, you are experiencing an ordinary reaction to thinking about electronic data as something with tangible properties like our stamps, coins, or audio CDs. The idea of a data collection *as* inherently incomplete is strange. I submit that we now do not have the words, classifications, or logical schema to discuss electronic information in the relatively simple forms in which it exists today. We are going to struggle like Sisyphus when virtual realities and cyberspace become commonplace.

This is not a trivial problem, nor will it be resolved quickly. Years if not decades will be required before terminology and catch-phrases crowd into everyday speech. One major stumbling block to management of the next generation of electronic information will be our inability to talk about the particular properties of information collections or data constructs. For example, what happens when one is confined to data that are selectively chosen from a particular type of source material? Is the accuracy of the data

The Collection

A Small Data Construct consists of structured records which is always incomplete. New records make the SDC larger. The collection consists of data from a single. object; for example, abstracts of articles.

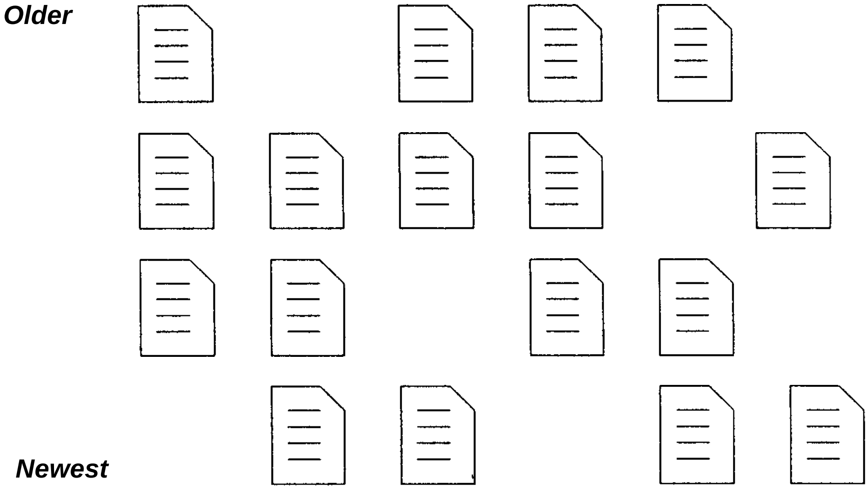


Figure 1.

different from a broader array of data from eclectic source material? Few consumers of electronic information come to grips with the issue of defining a stream of source material and then selecting only certain types of information from those sources. IS the resulting electronic file representative of the world as it is, or is it representative of the world as it appears when rules are applied to a specific stream of data by a controlling intelligence. This is not Big Brother; it is a database designer.

1 B. SELECTIVITY AND DATABASE CONTENT

Each time I raise the issue of selecting specific information for a database, I hear this question: "Do you mean that databases only take a few articles from each issue of a magazine?" I say, "Yes, that is exactly what I mean." Then I hear, "I had no idea . ..." Want proof? Do this: Pick an issue of Business Wee& or The Economist. Go online and limit your search to the citations from that specific issue of the periodical. Pass your search strategy against two or three databases that index, abstract, or present the full text of the magazine. Now compare the results of your searches with the original issue of the magazine. Who cited every article, story, letter to the editor, advertisement, and table? For those of you who did not take the test, the answer is no one.

What are the reasons for the selective approach to building a machine-readable file? I cannot cover all of the arguments, but I can cite a handful of representative ones:

- "Advertisements are not editorial content."-Exclusion by editorial policy.
- "Images, graphs, charts, and tables cannot be handled in our current production system or by our online vendors."-Exclusion by technological barrier.
- "The Letters to the Editor are too difficult to index and cannot be related to the article to which they refer because weeks if not months can separate the original article from the letter."-Exclusion by employee capability.
- "Our electronic feed is different in subtle ways from the printed version of the magazine because the publisher allows staff to make changes to the printing plates until the last possible minute. These changes do not appear in our electronic source."-Exclusion by source variation.
- "We have money to make a certain number of records each year. What do you want us to do: spend all our money on doing a few titles comprehensively or more titles by picking the most significant articles?"-Exclusion by financial barrier.

1C. VERIFICATION OF DATA

Callow journalists, consultants, and researchers know hat a "fact" is not "accurate" until verified. Database producers themselves suggest that a single query be passed across several databases. These are good instincts. I have not met anyone at trade shows or seminars who challenged me on the issue of selectivity in ABI/INFORM. Selectivity seems logical and a common sense solution to the problem of money, indexing difficulty, and so on. Unfortunately such gentle, indifferent thinking has some dramatic implications when we venture into the world of virtual-reality databases. Ready or not, we have to think about building electronic constructs that have unique properties.

The new machine-readable virtual-reality files will have an impact upon the way searchers and users of these data think about facts and accuracy. At this time, it is unlikely that someone can build machine-readable files that intentionally manipulate "reality." On the threshold of the new data constructs, we may want to think about the differences between old-style databases and those erected on cyberspace architectonics. When a database user believes that an electronic source can be accepted without critical evaluation or verification, that searcher will make decisions within the limits of the construct he relies upon. How can I argue with a person who has made a decision based upon "facts" obtained from an "authoritative" information source? The person with whom I disagree believes in the rightness of his position because he experiences first-hand the environment and saw the data as "real" and "valid." Experiences are his evidence.

1 D. SUMMARIZING THE CONCEPT OF A '/COLLECTION"

Let us come back to our original question: "How can a database resemble a collection of Ming ceramics or coins from Nero's reign?" Consider these similarities:

- A collection selects representative examples from a large universe of possibilities. It represents a subset of the larger reality from which it is drawn. Although a collection may include every example in the genus, we have other words to refer to a comprehensive set; for example, on/y and every. Only a few collectors can lay claim to comprehensive collections.
- Each item in the collection is indexed, cataloged, and arrayed so that it can be found using linear retrieval mechanisms (by date, alphabet, Boolean strings, etc.).
- One can add to a collection through time, but the contents of the collection are defined by understood or easily communicated "rules" or "assumptions" about what goes in; that is, the classification schemes are not complex; they are one-dimensional.
- In a database, each record has a structure. Mainframe DBMS once meant fixed-length fields. Even in micros, popular flat-file databases like Reflex, version 2.0 and Rapid File preserve this structure. New database implementations go well beyond a modest structural change from fixed-length to variable-length fields. New databases consist of a rich variety of data types and objects.

2A. HOMOGENEITY AND SMALL DATA CONSTRUCTS

Let me replace the term database with the phrase *small data construct* or SDC. The key point to remember about SDCs is that their records share a structure and well-defined common features like indexing or a specific type of information. For example, ABI/INFORM is a small data construct in this sense. The database has a sharply defined structure, indexing from a list of about 6,000 controlled terms, and an editorial policy that allows the database to gather and present information on a logically related group of subjects.

This database has about 300, 000 records and has been produced since 1971. One readily seen attribute of this database is that each record has visual similarity with any other record in the database. The homogeneity of appearance is evidence of the well-defined structure upon which the database is built. There are many databases that reflect these features. Most of the databases now available on Dialog Information Services and Maxwell Online exhibit some of these characteristics.

Small data construct does not refer to the number of records in a database, nor to the size of the individual records. An SDC echoes the idea of a collection or set of closely related objects. This type of data construct lends itself to the command-driven, question-and-answer interface: typing key words, building Boolean search strategies, or selecting choices from a menu that retrieves information. The best way to differentiate a small data construct from a database starting to evolve from a small data construct (SDC) to a large data construct (LDC) is to find files in which entrenched retrieval conventions don't work very well.

Boolean logic is almost foolproof with small data constructs. If someone wants information about a company, for example, the searcher selects a database from the timesharing company's library of files and enters the name of the company. The system responds with the number of items that match the query. Boolean operators allow the searcher to widen or narrow the set, including or excluding the records from the collection. Each subset of a small data construct is recognized to be a segment of the larger collection. It does not pretend to be a complete set of information on a particular query.

For a searcher familiar with the supermarket approach, a thorough investigation requires querying many different databases. With sequencing file queries, one can assemble comprehensive information about the topic of the search. There are exceptions, of course. If a searcher wants only the telephone number of the Komputerwerk computer company in Pittsburgh, PA, a single database like the Electronic Yellow Pages may well do the trick. Experienced searchers make an effort to doublecheck even the most basic facts.

Database producers tell their customers that other databases provide complementary information. Database producers recognize the limitations of their particular collections of information. It is naive to assume that a database has everything about a particular subject in one place. Even producers of massive small data constructs like the American Chemical Society's *Chemical* Abstracts surround their database with complementary files. Even Predicasts, obsessed as its managers are with Promt's new role as THE *mega-file,* state in their joint seminars with Dun's Marketing, Data Courier, and Dialog Information Services that other databases supplement Promt. The advertisements, however, demonstrate less modesty.

Collections always leave things out. Collections cannot be recursive. The omissions and shortcomings are most evident in databases that consist of text-only or number-only data. These are single-object databases, and they are the ones whose content is subject to cross-checking by the careful researcher.

2B. WHAT SMALL DATA CONSTRUCTS OMIT

What types of information do small data constructs omit? Before providing a selective list, let me remind the reader that the majority of electronic files or machine-readable databases today ignore visual and spoken information.

Few people would be willing to admit that an online inquiry is necessarily incomplete. Fewer still realize that the information retrieved may be presented in a slanted or biased way. An editorial policy and selective coverage of literature ensures these limitations.

However, it is quite difficult to obtain access to radio interviews with key business executives, video footage from television programs, or still photographs. There is a growing interest in these types of materials, and new small data constructs continue to make their way to market to serve the needs of some researchers. Burrell's, the clipping service based in New Jersey, has created a CD-ROM collection of transcripts of televised business news program. This product is a step in the right direction, but it is a small data construct and subject to the need for verification, multiple database querying, and the one-dimensional limitations of the small data construct itself.

Information omitted from the small data construct includes:

- non-searchable strings (recorded or synthesized voice)
- still images (photographs, line art, and technical and process color illustrations)
- motion photography
- video footage
- music.

We need to put aside he objections of traditional database producers at this point: "We are doing everything we can!" These professionals point to the technology barriers, the cost of modifying the existing database machinery, or the expense of getting rights and permissions.

There is no reliable figure for the amount of information that is not captured by the organizations creating small data constructs. When one considers the data transmitted on radio and television alone, we have access to a small percentage of the information disseminated each day. Database producers are quick to counter with the notion of significance. A database producer like University Microfilms (UMI) can say, 'We have the majority of the significant business and management literature." Who can argue? The only people who know are the editors of the database and a competitor like Predicasts, and both companies want customers to think ABI/INFORM and Promt have everything one needs in the way of business information.

Similar one-two arrangements exist throughout the small data construct world. What few searchers know is that both companies ignore far more information than they take. Neither UMI nor Predicasts, for example, cite every article in every publication they process in their database factories. Little wonder then that builders of small data constructs take a less-than-aggressive posture toward the types of data that fall outside their immediate production capabilities. If these companies cannot abstract every article in Business Week, how can they think about non-searchable strings in full-motion video and recorded speech?

3A. THE BOUNDARY-CROSSING CONSTRUCT

A powerful, relational microcomputer DBMS like Paradox 3.5 is simply incapable of handling different types of media like recorded speech. It is becoming increasingly evident that the inflexible database structures are being replaced by increasingly flexible designs. A newly developed microcomputer DBMS from Imaginetix (Pacific Grove, CA) accommodates text, images, and recorded speech. Search-and-retrieval software makes full use of the Microsoft Windows graphical environment. The differences between Paradox and the Imaginetix software is structural and conceptual. The gulf is wide.

Microcomputing horsepower, programming tools, and market needs force database producers to create more multidimensional files. The new DBMSs access a wide range of data types. The resulting databases are fundamentally new information constructs.

3B. CROSSING THE BOUNDARY BETWEEN SMALL AND LARGE DATA CONSTRUCTS

What is happening today is that the electronic collection (the traditional database like ABI/INFORM, for example) is evolving into a more different electronic construct. **UMI,** for instance, has attached images of the original full text articles to *some* of the abstracts in the ABI/INFORM Ondisc product the firm produces. The resulting database is no longer ABI/INFORM; it is a hybrid, created by grafting two complementary, yet unrelated, elements in one product. No one can deny that the fax images the product produces are gee-whiz. Nor can one doubt that the customer is not quite sure whether the product is a new type of database or an electronic form of microfilm.

Let me offer several other examples of change in the world of traditional databases:

- Images are gradually becoming available on commercial timesharing services like Dialog Information Services, but more rapidly on nontraditional services like underground bulletin board systems.

- CD-ROM and videodisc products routinely incorporate combinations of words, pictures, and sound.
- Software products like Danny Goodman's HyperCard and Owl's Guide allow users of microcomputers to combine different objects (text and pictures, for instance) in a flat-file database.
- Video capture and scanners are dropping in price. Their migration from the research lab to small-business office is underway.
- Display and output devices are getting more technically advanced and less expensive. They allow the user to view and print images in a farm roughly comparable to their paper analogues.
- Sound digitizers, including units adapted for game playing, bring high-quality stereophonic sound to IBM clones, not just machines designed for enhanced sound like the Macintosh.

Dozens of database producers are trying to develop new files that catch the wave of interest in mixing records with different types of data in a single file. Many of these efforts are primitive. The age of domination (1968-l 989) of small data constructs has ended.

3C. THE BOUNDARY CONDITION

Few database producers object to the rapid advances in technology that are being made today. They are, however, among the professionals least able to exploit these new technologies in their current electronic information products.

The 1990 National Online Meeting, the Special Libraries Association Summer 1990 conference, and the American Library Association meeting in June 1990 presented the attendee with a rich array of traditional electronic information products. At none of these shows was any database intended for commercial use that contained elements of virtual reality, cyberspace, or a large data construct. There were several clever "image" products from microfilm and software companies, but, in general, nothing showed the rich possibilities of the mixed-object environment.

There are companies, however, that are crossing the boundary between the small data construct and the large data construct. Recall Imaginetix. Bill and Linda Luther have devised a micro-based large data construct database tool. The implementation that the Luthers market allows a recruiter to capture an applicant's resume and embed recorded speech into the file. What this means is that the applicant can be interviewed, his responses recorded, verbal annotation made by the interviewer, and the database of objects reviewed by the potential employer. Obviously resumes of this type give the person screening candidates for a job a different sense of the applicant.

Another company-Voyager (Santa Clara, CA)-continues to offer optical ROM products that combine words, sound, and images for Macintosh computers. The Beethoven Ninth Symphony disc creates a self-contained world for the student of music. The score can be read; the music heard; a single instrument isolated; or the whole symphony can be played and stopped at any point for critical or reference information.

3D.RETRIEVAL OPTIONS

These products are starting to cross the boundary between the small data construct and the large data construct. The key driving power for the transition, of course, is the microcomputer's increasing power. As the local processing capability increases, it will be an inevitable extension of the small data construct to include the highly desirable objects like real-time video, full-color images, and sound. One can recognize boundary-crossing constructs because they will have several of these attributes:

- Multiple objects will be embedded in a traditional text (ASCII) database.
- Objects will be displayable on the screen simultaneously (which assumes that the objects are linked by indexing).
- The interface will be a combination of commands and graphics; that is, pull-down menus or keystrokes and mouse clicks.
- The various objects in the database will be searchable by type or in combination with the ASCII; that is, the data types will be held in separate files or tables and not be fully integrated into higher-level language objects like "envelopes," which can hold a collection of objects.
- User interfaces will be graphical. (In large data constructs, the user will be able to select an interface paradigm that suits his or her particular interests or skills; for example, a person may opt for an airplane control panel interface, not a mouse and keyboard. The control-deck paradigm makes it easier for the user to steer a search when flying over and through data.

3E. WHO WILL BUILD THE BOUNDARY FILES?

Traditional database producers are going to invent an assortment of boundary-crossing products. The revolutionary large data constructs will came from companies that watch the experiments closely. They will bring software, interface, and conceptual skills to their product design, which are seasoned by the pioneering efforts of others. These innovators are not likely to be trapped within the collection approach used in small data constructs.

This is not to say that a producer of a traditional database cannot successfully build a large data construct. I think that some striking innovations are likely to come from companies not now in the mainstream of electronic publishing. Another way to visualize the development time-line is to recognize that today's game players will be tomorrow's developers of large data constructs. These people see data as multiple objects. Abstracts will not set the next generation's imagination on fire. Tackling non-searchable strings-that is, recorded voice or synthesized speech, music, and full-motion video-may be a more interesting task. The boundary databases will carry the traditional information industry closer to cyberspace and virtual realities.

4A. THE LARGE DATA CONSTRUCT

If we look out five or six years, a new type of construct will become the basic information structure. The producers of these databases, however, are likely to be largely unknown in todays list of the who's who in electronic publishing.

This new type of database-see Figure 2 (p. 197), Two Database Architectonics-is emerging because traditional electronic information sources are useful but sharply constrained in heir ability to handle a variety of data objects. The builders of these machine-readable files are largely unaware that they are challenging the market domination of the small data construct that has held sway over the market since the late 1960s.

Nevertheless, the LDC is the database model of the future. It promises to have a profound impact on the way in which the users of the information process perceive the data within the construct.

4B. A WORKING DEFINITION

What's a large data construct? It is a database that contains multiple—that is, more than one object in a form other than ASCII text. For the purposes of this essay, let's create a large data construct. We define our sources as radio programs, video footage, photographs, news and analysis in text form, images of background documents, and appropriate indexing to these materials. We include graphical software to allow, for example, each 100th frame of a film to be shown in a 3×5-inch window on the screen. We have a speech accelerator module, which allows us to listen to normal speech in an accelerated form. We begin adding to each "record" charts and graphs, full text images for each citation, high-resolution photographs, and links among these objects (Figure 3, p. 198).

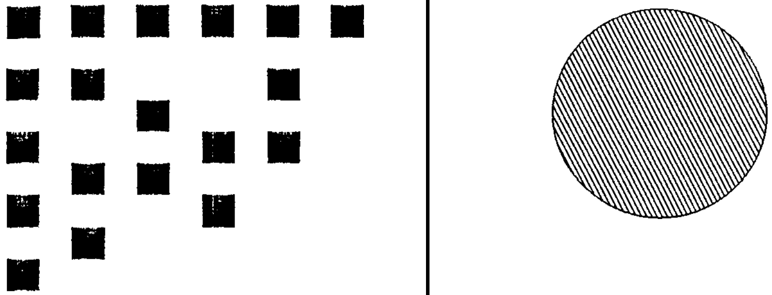
I suggest that providing these data in a three-dimensional, interactive space yields a large data construct. This hypothetical file was designed to contain multiple objects. It is not a traditional database with add-ons. We face no constraints about including information from different media. Access to this database is in a graphical, interactive reality roughly comparable to an arcade game in visual interest and dynamics. Searching the data is an exploration of information from a variety of media that retains as many of the attributes of "real" life as possible; that is, animation, collage, jumping from topic to topic, and so on.

Thus, although small data constructs can be enhanced with the addition of non-text objects, the addition of objects does not an LDC make. The large data construct is a machine-readable structure that has been optimized to contain searchable and non-searchable strings. It presents to the person querying the data a reality in which the ob-

Two Database Architectonics

Collection

Re-Presentation

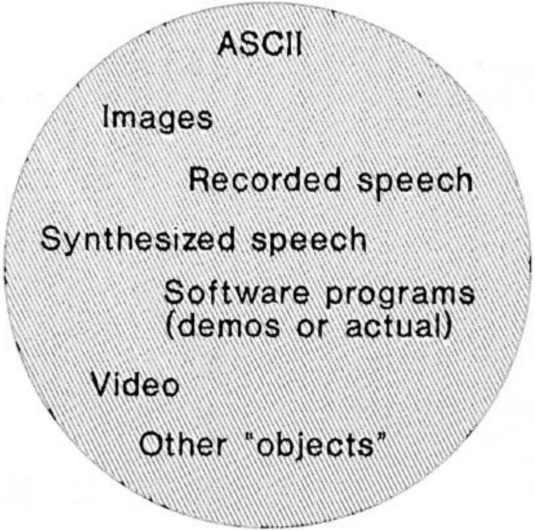


Must validate and cross check information.

Accept and apprehend information directly.

Figure 2.

Contents of a Re-Presentation



ASCII

Images

Recorded speech

Synthesized speech

Software programs (demos or actual)

Video

Other "objects"

Large Data Construct contains multiple objects.

Figure 3.

jects have the appearance of replicating (or at least suggesting) a tangible, believable world.

The large data construct requires a radically different type of software interface. The code gives the searcher seamless integration of data, a software interface suited to the user's needs, and advanced storage, output, and display technology. All combine to offer the user the appearance of reality. Who can say if the artificial reality is any more or less real than the actualities we experience with today's technology?

4C. SOME REASONS WHY LDCs ARE INEVITABLE

I have no doubt that large data constructs will be created, marketed, and used. Here's why:

They ore satisfying powerful needs. A large data construct is a user-need-driven, evolutionary step. Large data constructs will absorb small data constructs. There are numerous reasons for this. ASCII-only databases have reached a developmental end. Traditional online services and mast CD-ROM interfaces take the command-driven, query-response paradigm as far as it can go. To reach a market 100 times larger than the one that exists today, a new vehicle is needed. For the last three or four years, online systems have added features that do little to enhance the number of users who will spend significant amounts of money to interact with data. The radical escalation in marketing efforts designed to attract new customers signals the weaknesses of the traditional online model. Now marketers are trying to force the customers to use what they have rejected.

Experiencing data. The user-correctly or incorrectly-experiences the large data construct. When inside a large data construct, a user does not issue commands and get sets. The LDC is a re-presentation of reality, not a collection of separate items. A word of explanation may help the reader visualize the difference between the collection (small data construct) and the representation of reality (large data construct). Think of the interface to an arcade game. In the newer games, the data consist of multiple objects with which the user interacts. Games with a racecar interface let the customer steer, accelerate, and brake using the controls of a "real" car. The seat moves left and right, up and down to enhance the illusion that the customer is speeding around the racetrack. The number of interface paradigms is potentially large. Regardless of interface, the user willingly (or unwillingly) for the duration of the session suspends disbelief about the unreality of the game situation.

Going *inside data.* The user gets "into" the game. This is a function of the methodology of thought imposed by the need to manipulate and process multiple objects in real time. When a customer queries a large data construct, he is not willing or intellectually able to cease that interaction and query another data source unless that data source integrates seamlessly into the large data construct he accesses. Isn't something different taking place in the mind of the user of an LDC? I characterize the change as a 180-degree shift from the mindset or mental orientation a searcher has when querying a small data construct (traditional database).

Repeated exposure. In the LDC, a customer seeks repeated exposure to the construct. The more compelling the need to revisit the re-presentation, the more effective the LDC. That's why kids queue at arcades to play the hot games. Revisiting an SDC, in contrast, is driven by different needs. I search textual databases because I want to find and verify information. I can go online and do this because I am somewhat familiar with the command structure. Online and CD-ROM products are not, at this time, mass market products. What if I can get my data by driving a racecar to it? Will I continue to use the keyboard and commands? The LDC pushes today's market barriers back. It therefore follows that the more comprehensive, need-satisfying, and compelling the large data construct becomes, the less the customer's desire will be to cross-check the data or seek information elsewhere.

Self-reference. Large data constructs define their reality. I think of the large data construct as being an alternative information world-a self-contained information reality. The data in the large data construct define the electronic reality and provide information about that reality to the user. A large data construct-like an electronic game-contains its own internal logic, ethical system, and rules of accuracy.

Less emphasis upon reading. Large data constructs place less emphasis upon the processes of reading, comparing, and processing data. The small data construct forces the user to read, compare, and process in a sharply constrained way. Abstracts or full text articles have to be read. Images and sounds are apprehended, compared, and analyzed differently. Processing non-searchable strings like recorded sound and full-motion video bring other mental skills into play. The user of the LDC, however, may not listen to all available recordings pertinent to an experience. I assert that such an omission does not substantially alter the validity of the large data construct experience. Furthermore, the loss of certain data objects will not make the experience less real. Logical jumps or outright omissions are permissible. A video clip is assumed to be real and accurate even if the beginning, much of the middle, and the end are edited out.

What if an entire video clip is staged? The visitor to the virtual reality accepts that the large data construct is manipulative. A database need not be "true," "accurate," or "real" in the sense that database professionals now use the term. The large data construct is not subject to the rules of external logic. It is governed by its internal logic.

To summarize: The large data construct will contain multiple media that provide a wide range of information about the topic investigated. Much of the data will be absorbed by the user watching images or listening to sounds.

Whether comprehensive or not, the large data construct gives the user the impression that the data are complete within the environment the user accesses. The reason for this is that the large data construct defines a particular reality for a particular query. Familiarity with other large data constructs is tantamount to having first-hand knowledge of other realities. There is no reason to transfer experiences from one reality to another. Experience teaches the user that each representation is a self-standing, self-referential world.

No matter how we twist and turn, a re-presentation becomes a virtual reality. The data in a large data construct are "right"; they define themselves. Thus, the LDC achieves what in traditional databases is impossible: The LDC can both reference itself

and contain all other information relevant to the virtual reality. Consequently, it is far more difficult for the user to see the limitations of the information retrieved in the large data construct. In an LDC, the habitual user loses the ability to question the accuracy of the information.

4D. RE-PRESENTATION VERSUS COLLECTION

let me call attention to my use of the term re-presentation as a synonym for large data construct. The r-e-presentation gives the user of the large data construct an experience that is different from traditional library research and online research in a collection. The full extent of these differences, of course, is not yet known because we have an insufficient number of re-presentational databases which contain multiple objects, including video, images, full text, etc. Figure 4: Re-Presentation's Structure, (see p. 202) suggests that these next-generation databases will have complex, folded structures.

If all of this is disturbing to the reader, I urge him or her to recognize that large data constructs will create as many versions of reality, truth, accuracy, and data as there are database builders. The intellectual processes of capturing and internalizing the data are fundamentally different from the way in which people get information from a small data construct, databases will bring a dramatic change to our intellectual life.

Because large data constructs are electronic files, it will be possible to exchange these data environments. Clients can walk inside the data they have purchased from a consulting firm. When architects design a structure, the information about its layout, materials, and feel can be more easily communicated when the client and architect can share the mental model. When mental models about information are shared, thinking is more likely to be synchronized. The large data construct has a sense of place, of presence. It is qualitatively and intellectually different from analysis of data in a collection.

4E. EXPERIENCING THERE-PRESENTATION

The experience of data allows much more information to be transferred in a shorter amount of time. The conclusions drawn from this type of database search are valid, but they use thought processes that involve different types of information processing. which would you rather analyze: the experience of walking through a new building, or examining the printed sheets with specifications and drawings. The walk can take a matter of minutes; the analysis by traditional tools hours, days, weeks, or even months.

Let's end by recalling a trip to a natural history museum. Do you remember a display that shows animals in a stage setting that presents a frozen moment in time? The one I have in mind is beavers gnawing at trees. The larger museums spend thousands of dollars to make these dioramas or re-creations "real." Each blade of grass, the gaze of the animals, the foliage-all of the elements re-create a specific situation. My beavers

Re-Presentation's Structure

The space which a re-presentation occupies is complex.

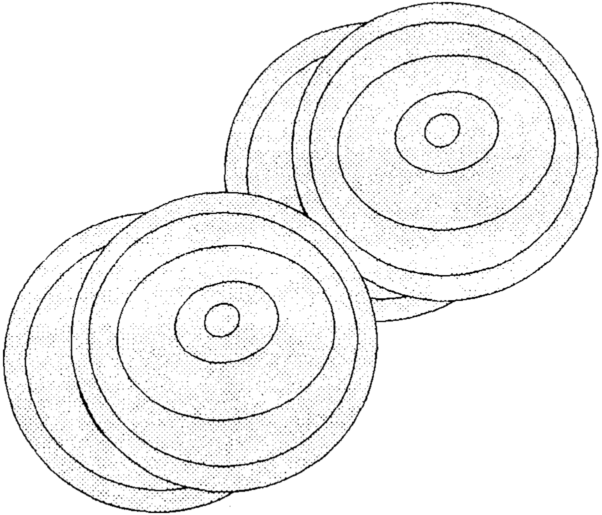


Figure 4.

are looking at a shrub from which a fox peers. To this day, I have never sought another diorama of beavers gnawing logs. The one I saw as a child did the trick for me. I don't need to see another beaver re-presentation. In fact, if I were to see live beavers in the wilderness or a re-presentation at another museum, I would probably doubt that they were "accurate." My mind has locked on to that first experience.

Such a mental fix is the strength and weakness of the LDC. We need to think about the implications of these dramatically new database architectonics. If we do not, how will we know when we have entered on and lost our ability to determine what is real and what is not, what is accurate and what is incorrect, and what is factually verifiable and what is not? I tried to duck the philosophical earlier. But here it is again, and, I fear, thinking about information in LDCs will raise the stakes in electronic publishing. One consolation: If you want the good old days, there will be a virtual reality for you in the not-too-distant future.

REFERENCES

Hall, Trish. (1990, July 8). "Virtual reality" takes its place in the real world. New *York Times* (National Edition), pp. 1, 12.

Lalonde, James E. (1990, July 8). World is virtually a reality: Cyberspace lab explores frontiers. *San Jose Mercury News,* pp. 1F, 7F.

Rheingold, Howard. (1990, Summer). Travels in virtual Reality. *Whole* Earth Review, No. 67, pp. *80-87.*