THE NEW INFORMATION PARADIGM THREAT OR OPPORTUNITY (OR BOTH)?

1996 Miles Conrad Memorial Lecture
Dr. Roger Summit, Dean, Founder and Chairman Emeritus, Dialog Information Services

ABSTRACT
Many of us remember the information explosion of the 1960s and the inadequacy of the tools we had to deal with it. The '70s saw the implementation of access technology using mainframe computers. This technology has been largely utilized by librarians and information specialists over the succeeding years.

The Internet, with its perceived vast hordes of freely available information, represents a new and perplexing paradigm. In this paper I hope to provide some perspective for understanding the current state of affairs.

In 1993, 1 attended my first high school class reunion -- the 45th. Although it only occupies four years of our lives, high school is an incredible period of exploration and socialization -- full of change and excitement -- which draws the students together in an atmosphere of challenge and reward. Not having visited my home town (Dearborn, Michigan) since the time I graduated, the reunion was a rich and moving experience.

We too have been drawn together through our professional association with each other over the past 10 to 20 years. Although I have only been away from you all for about four years, reuniting with so many friends and colleagues at this meeting is no less of a pleasurable experience. This is as it should be in that we have explored a period of technological change and challenge every bit as exciting as that we encountered as teenagers.

In several of the former Miles Conrad lectures, the issue of lack of cooperation amongst NFAIS members has been forcibly and probably appropriately raised. Nonetheless, we should not fail to realize that we have collectively accomplished a rather remarkable feat over this period of association, a feat that required an enormous amount of cooperation and mutual trust. Through our several efforts, a major portion of the world's published literature has been made accessible via computer terminals across the globe.

Once again, though, we are at the threshold of a new age. The development and evolution of the Internet over the past five years is a phenomenon unparalleled in my experience with computers and communications. Although originated many years ago by a government agency to facilitate communications between university and industrial researchers, the Internet has now exploded to a point that it will or is already a heavily impactive force in our lives and professional pursuits. With regard to the future, we only have to look at the market price commanded by some of the recent Internet-related IPOs (e.g., Netscape, Netcom, etc.) and the rapidity with which new developments are introduced (such as Hot Java, Alta Vista, Infoseek) to recognize that there is sufficient capital and technical talent focused on Internet activities to constitute critical mass.

What is the Internet?

Those of you who may not have had the curiosity and/or the time to surf and explore this new technology may find it all a bit mysterious and overwhelming. Let me attempt to provide an oversimplified but hopefully understandable overview of the Internet.

The Internet is a vast network of computers connected by high speed telecommunication lines. These computers and their software are called "servers." Users dial into the servers and using a so-called browser (e.g., Netscape), can leave messages on, or retrieve documents from, any other computer on the network. Each server, client, and document on the Internet has a unique address in a standard format. There are several search services, accessible though one's browser, which assist one in finding Internet documents of interest.

Use of the Internet is free to many users, or costs a nominal monthly fee for interconnection to a server. Incremental costs are normally zero (except with the so-called online services, such as America Online and CompuServe, which now offer their customers connection to the Internet).

Internet content consists of the masses of documents stored on these Internet-connected computers around the world. More and more the text is in a common SGML or HTML format which allows the documents to contain links to other related documents. For purposes of this paper, I will consider content as consisting of two types of Internet documents:

* Web pages

* Newsgroup postings on Usenet

Web pages are also known as Home Pages. Every Web page on the Net has an unique address (called an URL) which is typically of the form: http://www.pcd.stanford.edu.

Many Web pages are directories to files held and offered by the organization who maintains it. These files can be documents, databases, graphics, music, speech, movies, or combinations of these, all of which can be readily downloaded with an appropriate browser program such as Netscape.

Usenet Newsgroups are Internet discussion forums, along the lines of bulletin boards, to which users post documents, ask questions, and reply to others' postings. A favorite of mine is "rec.boats," wherein boating enthusiasts ask questions and post comments about their favorite topic.

There are many other services available on the Internet (e.g. electronic mail or e-mail) which are not discussed in this paper.

Internet pundits suggest that the whole nature of data processing is changing ... that networking and downloadable software will replace the boxes and boxes of application software we now buy ... that telecommunications and information are becoming nearly cost-free resources ... that decentralized data archives will replace periodicals and books ... that search engines will be so simple and powerful that one need merely feed them strings of words to perform effective searches ... that there is no need to index documents by virtue of the free-text search capability of the search engines... that a majority of commercial transactions will be conducted via the Internet... that the post office will become obsolete as personal and business communications evolve from stationary, envelopes, and stamps to electronic mail (e-mail) ... that we will now achieve the near-paperless office with significant conservation of forests ... that authors will self-publish and derive royalty fees directly ... the list goes on and on and on. Have you heard any of these predictions before?

Does this new technology threaten those of us in companies involved in traditional publishing, abstracting and indexing, and information retrieval services?

We must examine the Internet carefully before drawing firm conclusions. How large is the Internet compared to our databases? How effective are the search engines and services operating on the Internet?

In this Miles Conrad lecture I would like to examine this new paradigm (the Internet) in the context of the development and evolution of Dialog and the databases it hosts. Although history may not repeat itself exactly, the nouveau informationists seem to be discovering and/or relearning many of the same things we have faced over the past several decades. I will focus on search tools and the quantity and quality of content available to be searched.

The Past as a Prelude

In 1965 a meeting of senior scientists was held at the Lockheed Palo Alto Research Center to review a proposal to develop a computer-based information retrieval system--later to be called Dialog. Leon Fischer, a physicist, seemed agitated during the presentation.

The head of the lab, Dr. Jack Nash, however showed great interest in the idea of computers being able to assist in identifying documents of interest from a database of bibliographic citations. After all, the U.S. Air Force had specified "information retrieval" as a priority for authorizing independent research funding. Furthermore, there was a sense around the lab that it was usually easier to redo research than it was to determine if it had already been done (also, it was a lot more fun).

Perhaps it was the competition among the lab managers for the independent research budget, or perhaps it was truly his feeling that such an undertaking was frivolous that brought Dr. Fischer to hold out his approval of the project. His comment was, "I would never use a computer terminal to do a literature search; rather I'd call one of my esteemed colleagues and get the answer directly. It would be so much more convenient." (Ironically, in 1985 1 got a call from Dr. Fischer -- he had retired and was teaching at a local college -- to say how much he and his students enjoyed using Dialog, and how useful it had been to him as a teaching tool.)

Lockheed had just formed the Information Sciences Laboratory. The mission of the lab was to examine how third-generation hardware would affect computing in the information sciences. Third-generation hardware, typified by the IBM 360 computer, introduced mass, random-access storage, remotely controlled processing via telecommunications, and a time-sharing operation that allowed many people to utilize the computer at the same time. Following acceptance of my proposal, I was given responsibility for information retrieval.

Information retrieval was not a new concept, and there were several organizations doing tape searching utilizing second generation equipment. This approach was cumbersome. Although several searches could be run at once, the search specification could not be modified during the course of the search; and the output volume was unpredictable.

I felt that with third-generation technology we could design an interactive retrieval language that could overcome the inherent deficiencies of second generation equipment and could allow us to leap-frog over the existing services.

Early design objectives included the following:

* The user should be able to understand why results were retrieved based on the search statement, in order to be able to modify the search expression to improve the results.

* Index terms should display alphabetically near a candidate term, together with posting frequencies to assist the user in formulating a search expression

* Provision for nested Boolean expressions.

* Recursion -- the result of any search could be used as input in any subsequent search expression.

* The search engine had to be able to work with arbitrarily large files with no more than a logarithmic increase in processing time as files increased in size.

Nested Boolean expressions allowed the user flexibility in constructing search expressions from a simple, single-term search to a search of arbitrary complexity. Recursion allowed the user to factor a complex search into simple component searches and then to combine the results, as opposed to having to construct a single, highly-complex expression.

One of the early designs involved natural language (vs. Boolean) input, but it was discarded because of the frustration users felt in not knowing how to change the input to improve the output.

When we at Dialog decided to offer several databases, we defined a standard internal format into which the variety of formats used by different suppliers could be converted. This allowed us to have a single set of file load programs and also provided a product that allowed the user to learn a single search procedure that would apply to all databases on the system.

We further adopted a policy that even if "new is better," old is good. Consequently, we loaded all content available from suppliers and maintained even historical material online for as long as the database was offered.

Once into the commercial service, we intended to capture as much of the world's knowledge as was in computer-readable form and to offer a consistent retrieval language and consistent presentation. My dream was to be able to provide an answer to most any arbitrary interest or curiosity. Short-run profitability was secondary to product and service.

The evolution of retrieval service features is largely predictable -- today on the Internet as it was during our own development. If one service implements a good idea, the others are likely to follow as quickly as possible. A sample of these features from the past suggests the direction we can expect (or have already seen) from Internet services. Here are a few:

* Extend from carrying citation only to citation and abstract, and then the full text of documents

* Extend indexing from only selected fields, to the total document

* Extend databases from only the current portions, to the full archive

* Extend from searching on word occurrence in a document to proximity and/or nearness

* Extend from document-level to field- and sentence-level search specification

* Provide relevance ranking of output

* Provide for Boolean (if previously only allowing natural language search specifications) and vice versa

* Add an index display option

* Provide an SDI or current-awareness capability

* Provide for document delivery

Interestingly enough, each added feature complicates the user interface, so a difficult tradeoff must be made between complexity and desirability (although good design can sometimes mitigate some of the complexity).

The Present.
There are several search services on the Web. For example, The Bern Site (http://users.mwci.net/~bern/search.htm) provides an extensive list of Web search services:

Linkstar	Infoseek	Web Crawler
Yahoo	Telstra	Lycos
WWWWORM	Whole	Pronet
Nikos	Metasearch	Starting Point
Opentext	Galaxy	What's New Too?
Harvest	Apollo	WWW Yellow Pages
Nerdworld	Access One	Alta Vista
DejaNews	Webula	Bess

[Note to readers: Although Dr. Summit's original address had links to the above items, with the passage of time, many of these sites no longer are in operation. Consequently, NFAIS has chosen to display the table from Dr. Summit's address for the sake of accuracy but links are not provided]

Although several of these services use the same search engines, there is still considerable variability and some ambiguity among them with respect to how searches are specified, what variety of specification is allowed (e.g., free text vs. Boolean), the extent of the searchable corpus, special features, pricing (if any), stop words, updating frequency, back files, word proximity, provision of field specification, and in the definition of AND and OR operations (in at least one case AND in the conventional sense means OR, and vice versa).

Some services are directory services. Some cover only Web home pages, others only Usenet, still others cover both. There are still others that provide access to commercial databases. Some are free of charge, some require subscription, and some carry advertising. A common strategy is to become wildly popular with a free service and then to charge top dollar for advertisements interspersed randomly with retrieved pages. An examination of a sample of these services gives us a sense of this variety.

* Deja News (http://www.dejanews.com/) -- claims to be the "largest collection of indexed archived Usenet news anywhere!" But reading farther, you find that they exclude all Newsgroups beginning with alt*, soc* or talk* and all Newsgroups containing *binaries*. They archive up to one year in some cases and claim 4 gigabytes of searchable data. Although free to use now, they suggest there may be a fee later on.

* Excite (http://www.excite.com) -- claims to be the largest Web-site search service with over 50,000 Web pages and 10,000 news groups. Their purpose is to sell you search software by demonstrating its effectiveness on the Web. They claim that, "since in Excite there's no thesaurus or knowledge base to maintain, keeping your site up-to-date couldn't be easier -- just re-index as needed."

This last statement suggests to me that not only do you not need a thesaurus, but that if you want to use one, you're out of luck. The point about re-indexing suggests that there is not an index update facility and that the entire database needs to be rerun to generate index files each time anything is added.

* Infotrieve (http://www.infotrieve.com)-- offers Internet access to Medline on a flat-fee basis.

* Counterpoint (http://www.counterpoint.com) -- offers access to several government databases such as the Federal Register.

* NlightN (http://www.nlightn.com) -- is the online service of the Library Corporation and offers access to a number of public and commercial databases on an output-only charging basis.

* Alta Vista (http://altavista.digital.com)-- claims a very large index with access to all 10 billion words found in over 21 million Web pages. They also provide a full-text index of over 13,000 news groups updated in real time. In January of this year they were handling over two million URL requests per day.

* Infoseek (http://www.infoseek.com)-- claims to provide searching of all Web pages, Newsgroups, Usenet FAQs, and reviewed pages. They claim over 1 million Web pages, 10,000 Usenet groups (with 30 days' retention), and numerous other online resources. They have a free service and a fee service (which covers more of the Net).

* Open Text (http://www.opentext.com) -- claims a Web index which contains about 2.5 billion words of text. They apparently do not index Newsgroups.

* Yahoo! (http://www.yahoo.com)-- is a directory of the Web. It is a way of finding information by way of a subject hierarchy. If you think of the World Wide Web as a book (a really big, really disorganized book), Yahoo! is a table of contents. Yahoo! offers access to over 185,000 Web documents. Yahoo has recently joined with Open Text to offer searching as well

as directory look-up.

* Engineering Information Village (http://www.ei.org) -- offers commercial access to Compendex*Plus and the EiConnection which provides optional access to 150 other databases via Dialog.

* Knight-Ridder Information (http://www.dialog.com) -- offers access to commercial search services (Dialog and Datastar) and document delivery (SourceOne and UnCover).

Internet Size
One gets the sense that the size of the Internet is astronomical. I thought it would be interesting to compare the actual number of records available via Internet search services with those available on Dialog (many of which, as you know, are produced by NFAIS members).

Alta Vista claims to index all Web pages including the links within Web pages, which it estimates to be 21 million pages and 10 billion words. At 5 bytes per word, this leads to a total store of about 50 gigabytes for Web material. It is estimated that Usenet 'flow' is 150 to 200 mb per day. Alta Vista apparently archives 60 days' of most of the 13,000 to 15,000 Newsgroups, resulting in roughly 12.0 gb of additional material. The total for Alta Vista is thus 62 gb of text or 23 million documents, and represents the approximate total volume of public Web and Usenet information available on the Net.

Dialog, on the other hand carries an estimated 334 million records and 1.4 terabytes of retrievable information. Thus we can observe that the Dialog databases contain over 13 to 14 times the number of records and over 20 times the amount of accessible information that is publicly available on the Internet.

The following table summarizes this information and shows the comparative statistics for Alta Vista and Dialog. In addition it shows the approximate size of four major searchable segments of Dialog, each of which is several multiples of Internet.

Table I - Online Database Sizes

	Millions of Records	Billions of Words	Billions of Bytes
ALTA VISTA*	23.0	12.4	62.0
Web	21.0	10.0	50.0
Usenet(60 days)	2.0	2.4	12.0
DIALOG	334.5	280.0	1400.0
Science	167.1	139.8	699.0
Business	155.9	130.5	652.5
News	103.7	86.8	434.0
Text	76.6	64.1	320.6

*Alta Vista states "words" - I assume 5 characters/word on average.

The truly vast amount of information carried on Dialog-- much of which is supplied by NFAIS members -- suggests to me a continuing role for abstracting and indexing organizations and information retrieval services. Information retrieval is a process of successive screenings, that is of winnowings down. First look at titles; from selected titles look at abstracts; from selected abstracts, download or order the full document. Searching only the full text of a very large collection of full-text documents is not only cumbersome and slow, but is likely to yield a large result that will need to be examined sequentially.

Quality of Internet information.
The statistics still beg an important qualitative question. What is on the Web (and in Newsgroups) and what is in commercial retrieval services? Although the Internet contains some very valuable and interesting information, much of it is entertainment related -- or communication oriented -- and only

some of it is moderated or refereed. Tim Miller addresses this point in the February 1996 issue of Information World Review:

> But the information nirvana is turning into information purgatory. The Web has become a Gutenberg on steroids, spawning millions of pages of new content a year, much of it of uncertain lineage and some of dubious quality. A great deal of end-user time at work is clearly spent, according to observations of directory providers, on entertainment or titillation. And when they do want information related to their professional objectives, users can easily spend hours of fruitless browsing with numerous opportunities for distraction. The scope for wasted time, money and opportunity is vast.

Carlos Cuadra used to say the way to succeed financially in the information business is to find a database of pornography with low unit storage cost that could command a high unit price. Many Internet services have found the secret formula, except for the high unit price. Oh well, with popular material there's always a trade-off between price and volume.

If this theme sounds familiar, remember the French Teletel service. Telephony [vol. 215, no. 6, pp. 24-5, Aug 8, 1988] indicated that critics of the system (French Teletel) cite a lack of local intelligence and memory storage, slow performance, and an overabundance of pornographic services (my italics). A more recent article indicates Teletel is cutting prices due to competition from the Internet. [Competition and too much pornography in France? Will wonders never cease?]

Searching the Internet

How can one make the decision of whether to search the Internet or a commercial service? A definitive answer is difficult to give, but after one uses both for a while on a variety of questions, you begin to get a sense of what types of questions find productive answers on which services. My guess is that someone far more scientific and methodical than I will do a comparative study of overall retrieval effectiveness -- considering the combined power of the access language and the content -- and give us some definitive answers.

In the meantime I would like to report an interesting and enjoyable project I was invited to participate in last month. If you are a parent or a librarian in the San Francisco Bay Area you are probably aware of the annual Millard Fillmore Trivia Hunt. Teams of students from several high schools in the area are given a list of some 30 to 40 questions on Friday evening. They have until Sunday to come up with the answers. Each answer must cite two independent sources for the information. They can use any resources they can muster including local libraries (and librarians), teachers, parents, and whatever.

This year the students made a lot of use of the Internet. Saturday afternoon -- about half way through the competition -- I got a call from a teacher at Woodside High who mentioned they were really stuck on one of the questions, and could I help them out? The question was: "Mat is the name of the jazz musician who died of pneumonia in 1943 while riding the Santa Fe Railroad?" The answer: Tom Waller (better know as 'Fats'). The answer was in the January 3 issue of the Arizona Republic. The kids were delighted.

Late Sunday afternoon near the end of the contest, I got another call, this time from a student who asked if he could send me a few more of their tough questions. I tried the questions both on the Internet (variously using Alta Vista, Excite and Open Text) and Dialog. On Dialog I tended to use File 411, the all files index, to determine which databases to search, then searched the databases for the answers. Final score: Internet found 3, missed 5; Dialog found 6 missed 2.

If I were asked to provide a guideline on which to search, it would be the following: If you can afford the time, do not need to be exhaustive, and the topic is contemporary -- try the Internet. Otherwise use a commercial online retrieval service. On anything important, I would use both.

Publishers and the Internet

Just as Internet search services are not likely to replace commercial retrieval services, online publishing is not likely to replace the traditional publisher. Both can coexist and even flourish on the Internet. A recent example will illustrate the point.

Steve Summit (my nephew) maintains the FAQ (Frequently Asked Questions) group for the C programming language on the Internet. He was approached last year by Addison-Wesley about publishing the FAQ in book form. Steve, in common with many serious Internet enthusiasts, was quite concerned with possible commercial conflict with what has been for him a devoted and benevolent pursuit. Now that the book is published [LCCN: 95039682 //r96; C Programming FAQs; frequently asked questions / Steve Summit], I asked him how he felt about the experience. Our conversation (via e-mail) is summarized as follows:

Q:
Overall were you pleased or displeased with your decision to publish?

A:
My initial concerns over copyright encumbrance of the formerly freely-redistributable FAQ list evaporated as Addison-Wesley bent over backwards to ensure that free redistribution (which they view not as competition but as free publicity) would continue. Their support certainly benefitted me, presumably benefitted them, and benefitted the Net as well, because their ponsorship enabled me to make needed improvements to the list, which I was running out of incentives to do on my own. They set a fine example that I hope they and other publishers will be able to maintain.

Q:
I can imagine that commercial availability could even increase the amount of Internet downloads due to publicity.

A:
Certainly. And the senior editor on my project views free Internet copies not, in the old-guard paranoid view, as lost sales, but rather as free publicity. I was flabbergasted when he said this; I (in my own cynicism) had assumed that he couldn't possibly have this outlook, and that I'd have to fight tooth and nail to convince him, and probably not succeed. [As it turned out] Addison Wesley had offered me far more than I'd ever have tried to ask for.

Q:
If you would care to share it, I am interested in the part of your contract with the publisher relating to electronic rights you/they retain.

A:
... here's a description of one of the relevant paragraphs:
The Publisher hereby grants you the non-exclusive right to publish an electronic version of the book. In the event that you initiate a program to sell an electronic version of the work, you agree to pay us 20 percent of net revenue received from sales of the work. You and the Publisher will not enter into any electronic distribution agreement without prior approval, which will not be unreasonably withheld .... There was also a clause in their standard contract - a very fair and forward-thinking one, I thought -- that said that royalties for modes of publication not mentioned (or even thought of yet) would be paid, by agreement, at least at the then prevailing industry rate.

With all of the controversy between the National Writers' Guild and primary publishers, the above would seem to supply a reasonable model at least for future contracts.

Concluding Remarks
Nothing in my earlier remarks is meant to suggest that the Internet is a passing fad or that it will give way to more traditional means of information transfer. If it contains only a fraction of the amount of information of commercial information services, and if it contains so much informal and unmoderated information, why is the Internet growing so rapidly? and why is it so attractive to developers and investors? To my mind the answers lie in several successive and critical steps that have occurred in the evolution of the Internet and the highly important features it offers as follows:

* Free or virtually free telecommunications, first to ARPA and now Internet users, has encouraged the open interchange of ideas via e-mail and led to the development of tools for exchanging documents, graphics, and computer programs.

* Vice President Gore's hyping of the Information Highway carried a lot of media publicity to the public resulting in further growth.

* The decision by popular online services (such as CompuServe and America Online) to connect their already established customer bases to the Internet, expanded the market.

* The international scope of the network provided the ability to communicate with people worldwide.

* A philosophy of free exchange, which has led to volumes of contributed documents, images, music, and software, allowed the ordinary user to easily and conveniently contact and obtain information from experts (real people) on any variety of topics.

* For the first time, a medium of information exchange is largely independent of platform (computer type).

* The free distribution of browser software first Mosaic and then Netscape, has promoted access.

* The large potential market has encouraged development of products, which could not be justified utilizing traditional advertising and distribution channels.

* Because the Internet does not rely on face-to-face contact, it represents a unification medium independent of race, sex, age, title, or disability. Nobody knows if you are ugly, educated, short, or wealthy. Your discourse is judged on its own merits. Many of you have seen the New Yorker cartoon showing a dog at a terminal who says something like, "It's great, on the Internet, nobody knows you're a dog."

Aside from the prohibitions and ambiguities of the Telecommunications Act of 1996, are there any threats to continued explosive development for the Net? In his editorial in the March issue of Internet World, Michael Neubarth expresses concern that in its transition to a major market, the Internet is losing the sense of community that promoted past cooperative efforts. He goes on to say that if companies like Microsoft, Netscape, and Sun in their fierce battle for market share gain dominance they could poison the wellspring of openness and innovation that has been enjoyed in the past. He concludes with a statement that it would be poetic justice if the Internet, built to withstand nuclear war, also could thwart commercial attempts to dominate it.

If the Internet can withstand these efforts to control and dominate it, it can provide a forum for ease of entry for new innovations, a medium which can expose traditional products to much wider markets, and an environment wherein success or failure will depend upon real quality, not hype.

If this isn't exciting, I don't know what is.

As the former President and Chief Executive Officer of Dialog Information Services, Inc., Dr. Summit's pioneering work on development of the DIALOG system began in 1962 at the Lockheed Corporation. He was designer and project manager for the first-of-its-kind online information retrieval system in the early 1960s and continued in various executive capacities until he retired at the end of 1991. He has received many awards, and he has held several positions in professional associations and on advisory boards. He has presented and/or published over100 papers and journal articles. Currently, he is a Library Commissioner for his town of residence and serves on the boards of several companies and nonprofit organizations. His education includes a B.A. in Psychology (1952), an M.B.A. (1957), and a Ph.D. in Management Science (1965), all from Stanford University.