# Information Analysis in the Net

## The Interspace of the Twenty-First Century

Bruce R. Schatz

This is going to be an odd discussion, and it is not just because I am kind of an odd person—although that is true—and not just because it was done at the last minute, but because what I am really going to do is talk about the future. The easiest way to talk about the future is to look for ways to do prediction, and the best way that I know to do that is to look at what is going on in the research area—i.e., to look at big research systems that are trying to show what things will be like in the distant future. "A long time" used to be fifteen or twenty years; these days, a long time is five or ten years because the world is just moving so much faster, but using whatever the research systems are doing now to predict what the world will be like in the future is still a good method. So I will be going through much technology very quickly, and you should not scrutinize any of the details but just see what the flow is and what the main idea is.

First, the discussion will center around how prediction of technology trends has gone in the past. So the discussion will be about the evolution of the Net and where things are going, just very briefly, as well as about the evolution of the Net and where things are going by providing a historical example. In fact, I am going to talk briefly about the "Telesophy" system and how the predictions worked in that case, then talk about what present research systems are like to provide an impression of what they will be like in the future.

So, what is the state of the Net? Well, right now, even with this great excitement about 50 million users, the issue is primarily on access. Presently, we are really doing access, and if you look at something like Mosaic and are feeling very excited about it, remember that it is just suped-up FTP—i.e., it is fetching documents. It might be very streamlined but is not actually finding anything. You point to something and are able to get that very transparently. What you will see in the next wave is more like

- **The Present: Access**
  **—The Net fetches documents**
- **The Future: Organization**
  **—The Net searches repositories**
- **The Millennium: Analysis**
  **—The Net correlates information**

**From the Internet (data transmission) to the Interspace (information manipulation)**

Figure 1. Evolution of the Net

organization, which is what you are used to, being able to do a real search—like what online retrieval systems have done in the commercial market, like Dialog, for a long time. To describe organization, you will usually hear phrases like "searching repositories" which is what the Digital Library projects are doing.

Then consider the further future. The reason I am saying millennium is because, if you think about it, the next millennium is only two years away, in the year 2000. Things will be very different then, and what will actually happen is that ordinary people will be able to solve real information problems themselves, and you will see more about correlating information than doing searching. So the second part of this discussion is about what the future is going to be like—how you really are going to be able to do analysis. This prediction uses the coming research technology as a future projection. As a grand statement, you can say that we all will be moving from something like the Internet to something like the Interspace. Much more will be said about what the Interspace is and why you would want to call it that rather than call it the Internet.

- **The Internet as File Transfer**
- **TCP/IP and FTP**
- **Gopher and WAIS**
- **Mosaic and WWW**

**Fetch in the Net**

Figure 2. The Present: Access

So here we are in the present. Access is basically file transfers (TCP/IP and FTP, Gopher and WAIS, Mosaic, and WWW). Now Gopher and

Mosaic are things that did not even exist five years ago, and today there is, honestly, 50 million copies of Web browsers of different kinds. People are using them everyday to fetch transparently things that, in the past, people were unable to fetch, that only were available to the cognoscenti, so that is really fetching.

- **The Internet as Information Repository**
- **SGML and document structure**
- **Metadata and Depositing**
- **A & I: Abstracting & Indexing**

## Search in the Net

Figure 3. The Future: Organization

Then in Figure 3 is an outline of the kind of thing you will see in, say, the next two or three years, and you are already starting to see with Internet search services. What you will observe is that you will be able to do a search, and there will be repositories where you can actually put information. There are search engines that search in different ways and there are higher level directories that you will be able to use to find things around the Net. In the research domain, these are things like the digital library projects that were discussed before, but there is also very large commercial activity in this area. Some of the issues again are: how you actually structure the documents (SGML is a tagging scheme where you indicate the fine-grain parts), how you record what is called metadata in the technical sense (you probably think of these fields as the bibliographic citations on the outside of the document), and how you really go about doing indexing and such.

- **The Interspace as Concept Grouping**
- **Community Repositories**
- **Computer-Assisted Indexing**
- **Vocabulary-Switched Retrieval**

## Correlation in the Net

Figure 4. The Millennium: Analysis

So, the majority of this discussion is going to be about the distant future and not the immediate future. To provide an indication of what you should be thinking about—that which is going to happen in your

working lifetime—it will be possible to move beyond merely searching documents so that you are actually handling concepts and manipulating them. You will have repositories for groups and collections too small and informal to be handled by professional indexers, not like something for electrical engineers but down to the fine-grain community level—where a community might be ten people locally that have a karate club or a hundred people around the country that have a karate club.

In fact, this goes back to the plenary talk done earlier that mentioned a beautiful quote from a graduate student that said: "The Net is not about information, it's about community, it's about sharing." That is going to be very true. The history of what the electronic medium for which the Net is used, all the way back to the videotex stage, shows that what people really want is to swap information and store particular things they care about, not access big centralized collections. So there will be much capability for doing that swapping and, in order to do that, you need an underlying infrastructure which will do much more than is possible now and well beyond full-text search. So you will see things like support for domain experts who don't know enough about classification to enable them to do effective indexing. You will see support for being able to switch vocabulary across subject domains, and I'll talk at length about what the technology would be like for that, because there are already instances in the research area of being able to show that functionality.

- **1986 Telesophy prototype at Bellcore**
- **1989 Schatz advisor at NCSA**
- **1991 WCS prototype at Arizona**
- **1993 Mosaic developed at NCSA**
- **1994 1M users of Mosaic on the Net**
- **1995 Netscape worth $5B; 10M users**
- **1996 online services (AOL); 50M users**

**research prototype → mass commercial**

Figure 5. Information System Timeline

The discussion will now turn to the historical—i.e., how one might predict the future by using a set of examples that were developed by myself. I also know the subjects very well, and these examples are illustrative of what the prediction process is like. There is a ten-year period—that used to be a fifteen to twenty year period—from when you had a working prototype of something in the lab to when it was a $1 billion business and millions of people were using it. These days the time line is much shorter.

The period of time used here was ten years, and many people predicted it would be twenty or thirty and, in the future, it might even be five or less. So, this is basically a time line. Much will be said about what the telesophy prototype was so that one can contrast it ten years later to what the Net has become because, for example, you are all probably very familiar with what Web browsers do.

In 1989, I became the Scientific Advisor at NCSA for information systems. Nobody knew what that was. They were doing very well with NCSA Telnet, but they had this idea, because of a very progressive director named Larry Smarr, that someday the Net would be a great source of valuable information for the scientific community—i.e., there was this crazy fellow (me) who had done a big project on something revolutionary, so the powers that be thought I could show up occasionally and inspire the troops. Well, I had in the meantime moved on from Bellcore to the University of Arizona and, in 1991, I produced this Worm Community System that I will say a few things about because it is a useful historical analogy. Then, in 1993, after several attempts to reproduce telesophy on more of a mass scale, good enough underlying technology, which in this case was the World Wide Web, finally became available, and Mosaic was developed. Mosaic was a relatively small effort at first, then developed into something involving about ten to fifteen full-time programmers.

But then the world exploded. Look at the time line; this is what surprised everybody: 1 million users the next year, 10 million users the next year, now there is a company that evolved from it worth $2 billion in essentially eighteen months—this was Netscape. Most of the projections for 1996, which you see is only ten years away from Telesophy, was that there will be 50 million users/online searchers on the Net. Thus this ten year time period is very striking because it is no longer an esoteric subject anymore. So, now a little about what things were like ten years ago, and you will see that they actually were fairly good predictors of what was found ten years later, so then when the grand vision of things for the future is

- **vision of transparent knowledge manipulation**
- **multimedia information retrieval**
- **wide range of information sources**
- **search across distributed repositories**
- **grouping and sharing basic features**
- **good performance and scalable architecture**

**Distributed Access and Organization**

Figure 6. Telesophy System

explained, you might perhaps believe that there is some predictive value in what research systems are like.

So, what do you really want? You want something that has all the world's information in it, something you can browse around in, and that is all interconnected—but that sounds like science fiction. So, what you actually have now are things like this Telesophy system that will be described shortly. How do you get from here to there? Well, you have to lay fibers, you have to harden the software, and you have to have more powerful machines. What happened is that the technology curves were much faster than anyone predicted—e.g., personal computers became cheaper much more quickly than people predicted, and network speeds became faster also much more quickly than people predicted. The software did not really get any better, but that is always true with software.

Telesophy was to be the universal system between all the world's knowledge and all the world's people. People are putting things in and getting things out, so that you can sit on the far end with your portal into information space, go out over a switched network and get all the world's knowledge, different types and different locations. So, basically, you can get anything from anywhere, and the system underneath hides everything. Telesophy is similar to telephony—"tele" means "at a distance" and "sophy" is "like wisdom or knowledge." Just as the telephone hides all the sound from places, the telesophy portal hides that you are getting all this knowledge from other places and does not tell you what happens underneath.

Well, there was this grand vision in a lengthy report I wrote about technological feasibility. I also built a prototype, and what impressed people most was that the Telesophy prototype actually demonstrated the vision with real technology and a real architecture. There the prototype was. It did multimedia information retrieval across real networks, and it had a wide range of different sources. You could actually put repositories in different places (what would now be called repositories) and search across all of them. It had ways of saving what you found, the results of searches, and storing these for use later. It ran pretty quickly and it allegedly scaled up.

The prototype running in 1986 had about twenty sources that ranged from messages like wire services to citations like Inspec and MEDLINE to full text like magazine articles and movie reviews to library catalogs to a sampling of multimedia things like line graphics and color pictures and motion videos. You could sit at a workstation and search across all these sources for broad terms like "fiber." The system would then search all the sources (they were all carefully indexed), bring back in real-time the matching items from each source, and you could manipulate them. First, the system would show a one-line description then, if you wanted more details, it would pull up a picture or the full text of an article. If there was a

- **Community:** 60 Bellcore users, 40 external sites
- **Environment:** workstation network (C/Unix on Suns)
- **Software:** local object system, remote searching system
- **Type Transparency:** text, graphics, images, video
- **Location Transparency:** 20 servers (TCP/IP)
- **Scale Transparency:** 300K units in information space
- **Sources:** messages (news, newswire), citations (Inspec, Medline), full-text (magazines, reviews), catalogs (library, memos), pictorial (images, videos)
- **Networks:** universal access within Bellcore Internet response feels like library browsing (building & WAN)

Figure 7. Telesophy Prototype

link in that article to something else, you could just push a button, and it would jump to that link automatically.

You could also make new information out of old. While you were searching, you could pull something from here and something from there and something from there, then combine them to create a new piece of information with some classification notation for later retrieval. So, for example, you could save a set of documents or pictures that you retrieved. It worked the same with pictures or videos. What I was going to demonstrate with the 35mm slides—which were too dim because of their age—was a slide of me sitting at my desk at Bellcore, searching all these sources, then pulling the camera back to show "yes, this is really my desk and it did have color pictures and it did have video, and it did have this session searching, and it was actually working."

The prototype was used every day for several years, and there was a limited number of other people besides myself that used it, but the problem was that it was sort of a "hero" experiment. It had fairly expensive workstation equipment, costing about $30,000. It relied strongly on having a fast local network, which was very uncommon at that time. It was very hard to collect enough data in the right formats to actually search it, and even now when you try to run experiments, like in the Digital Library project, that still tends to be true. The reason that the prototype was impressive was that one could run the technology curves up and say "yes, if there really was a megabit fiber network everywhere and you had a personal computer that was like a Sun workstation, then you could do the same thing from home." And the reason that this kind of technology took ten years to hit the mass market instead of twenty or thirty is that no one predicted how fast the hardware curves would go down.

The telesophy system was thus a good predictor of what the future would hold ten years later. In functionality it was a superset of what Web browsers are now, and it was actually fairly close to what Web browsers will be in one or two years, because it had ways of adding personal materials. The collaboration facilities are just starting to enter the Internet now, but we will be there fairly soon. Telesophy also had good search capabilities across multiple sources, at least straight full-text search, and this is just beginning to become the standard on the Net.

I regret to say that what an earlier presenter said was exactly true. Bellcore felt that the future of electronic information was video-on-demand, so they thought telesophy was an interesting high-runner project, and they put money into it for a couple of years, but when it became a question of fishing or cutting bait, they decided to cut bait. Thus they chose not to invest a substantial amount of money into this and, in fact, they also passed up, despite some serious discussion, a chance to patent the concept of information spaces because it was felt that a software patent was not defensible, and it was not going to be an important enough area. I have since had discussions that indicated they could have owned the Web—i.e., the Web would have been an infringement of their patent. Sorry to say, that is just one of the corporate decisions. It would not have made me personally rich, and may have been just as well since Bellcore probably would have clamped down on its propagation, and thus the Web would not have spread as quickly. Such stories often happen in the history of technology.

In the model of a telesophy system, there was this thing called an information space, with real data down at the bottom, and these little packages, called information units, which were uniform across all the data in all the sources. Information units were object packages that had uniform formats that enabled the system to search across everything or group across anything. After a search, the filtered results could be bundled into a single information unit that could be displayed and searched (sort of a knowledge region) even though it was actually a group of items of different types in physically different places.

So, in summary, what the Telesophy system showed was that one could really do transparency of type and location. It did not show scale in billions very well, but it certainly showed scale in millions. There were about a million—well, about 300,000 to 400,000—items in the whole space, and it was tuned and fast so that one would really get a one second response for a search and a half-second response when you clicked and tried to follow a link. So the prototype also showed you could really do things fast like you were wandering through a library. Much of the technology, much of the implementation effort, was attempting to make browsing a worldwide electronic library at least as functional as a physical one. The prototype then tried to show some grander things which were more technical.

| Access | → | Organization | → | Analysis |
|---|---|---|---|---|
| web fetch | | library search | | space<br>correlate |

**so what's the 1996 research system that will be the mass service of 2006?**

**cross-correlation from multiple sources**
**generic community system a la WCS**
**interconnection of spaces above networks**

Figure 8. Towards the Interspace

That is what happened in the past, and about ten years ago it was clear that Net browsing could be done and would be big and grand and many people would use it. Big and grand turned out to be ten years. So now the discussion will turn to what is going to happen ten years from now. The real question is, in the twenty-first century, what are there going to be 50 million or a 100 million copies of?

My hypothesis, as you can probably tell, is that it is going to be whatever it is possible to be included in a large research system. So, the remaining discussion will be about what can actually be done if you do a grand hero experiment and then extrapolate for yourself with whatever kind of historical analogies you like as to whether that will happen and, if so, when? My belief is that it will be a $1 billion business in the early twenty-first century. And what is *it*? It is not Web fetching, which is just straight access, it is not library search, which is what you will see in the next few years when you can put up a big collection and actually search it. It is going to be correlation, analysis, coming in with a real problem and being able to look through numerous different sources and saying, "this thing here and this thing here combined in this certain way solves my problem." Some analogies of that from several projects will be provided in an attempt to give some concrete feeling, and then the technology will be discussed.

So I now will talk about cross-correlation, generic community systems, and spaces not networks. Those probably do not mean very much right now, but an attempt will be made to give enough examples so that you can get a feeling for what those concepts might actually mean.

First of all, a little bit about WCS (Worm Community System). This is what I personally was working on in the five years that Mosaic was starting up. What that tried to do is, essentially, make a real telesophy system in a small area of molecular biology and see what was really involved. It was trying to build an electronic scientific community that had data and lit-

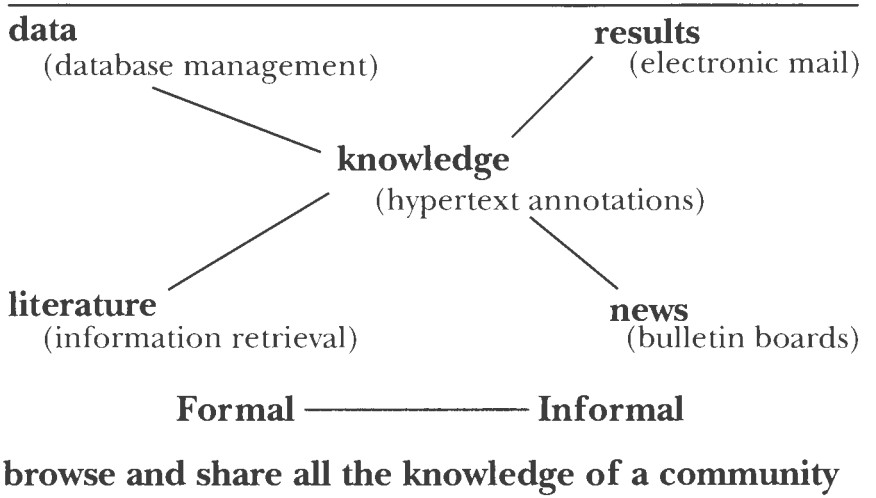


Figure 9. Community System

- **WCS Information:**

| | |
|---|---|
| **Literature** | **Biosis, Medline, newsletter, meetings** |
| **Data** | **Genes, Maps, Sequences, strains, people** |

- **WCS Functionality**

| | |
|---|---|
| **Browsing** | **search, navigation** |
| **Filtering** | **selection, analysis** |
| **Sharing** | **linking, publishing** |

- **WCS: 250 users at 50 labs across the Internet**

Figure 10. Worm Community System

erature both informal and formal. So it had real databases in biology and real literature. It also had bulletin boards—i.e., informal information, like community newsletters and meetings. You would sit there in this single space and could search across everything to select desired information. Then you could follow very fine grain links so, if there was mention of a particular gene that you were looking for in an article, you could jump right to the corresponding item in the related database. You could also take a display of a database item and pass it into another program.

WCS tried to handle all the knowledge in this small community and wanted to be able to manipulate it, both taking it out and putting it in and passing it into programs. As mentioned earlier, it had basically all these functions. You could browse, search, navigate, follow links, or select part of a map or gene description and pass it into another program. In addition, you could, since it was a symmetric system, add anything that was supported within the system. You could add your own gene descriptions, make a link between your gene and this other gene, or do a submission on the fly to one of the main databases.

So, essentially, information needs were handled within this single environment, and what happened over a five year period was that a working system was built and evolved. There were several hundred different users and about fifty different labs that were actually using this system, at least on a test basis. They used it mostly for information retrieval, but they also did some resource sharing. It did span all the connections and had very fine-grained editorial control. You could actually publish things. You could also keep them private for a while then move them out to the next level database. So, it also tried to capture the complete publishing cycle.

Then basically what happened is what usually happens to research systems, which is that the good ideas were absorbed in a more popularized fashion into other (low end) systems that were trying to appeal more to the masses, and the research system itself disappeared. So what happened in this particular case is the genome projects took over much of the nice graphical displays with the link following, and Mosaic and the Web browsers took over the fetching part across the Internet, and the Worm System disappeared. But it showed what was possible to do in handling all the knowledge of a small-size community.

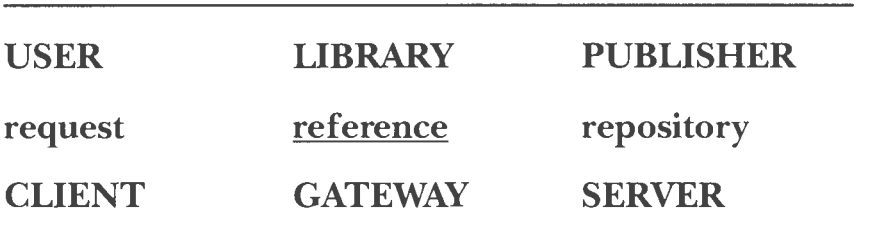


Figure 11. Distributed Library Model

So here is the second of two different metaphor types to explain what Interspace should be. The first was taking a whole community and handling all the information in it a la WCS, and this second one is sort of what librarians really do—real libraries.

If you look at the digital library project or a physical library, usually you think of it as "here's a big repository and here's the user and they want to do a search in there." Well, that really is not what librarians do. What librarians do is, they know numerous sources and have a huge library with many books and from many sources, and there are many sources that they know are not physically in the library. Mostly what they are doing is serving as a reference, as Figure 11 shows in the middle as a gateway or reference. They are trying to solve a particular information problem for a user by routing them here and letting a user look at that, and routing them there and letting the user look at that, so that they are going through many different sources in a reference session trying to solve a problem by correlating the parts. Well, digital libraries do not do that. They just do a straight search. But suppose that reference was now the most important thing. Suppose you could search and access and you can do organization, then you would want to do correlation.

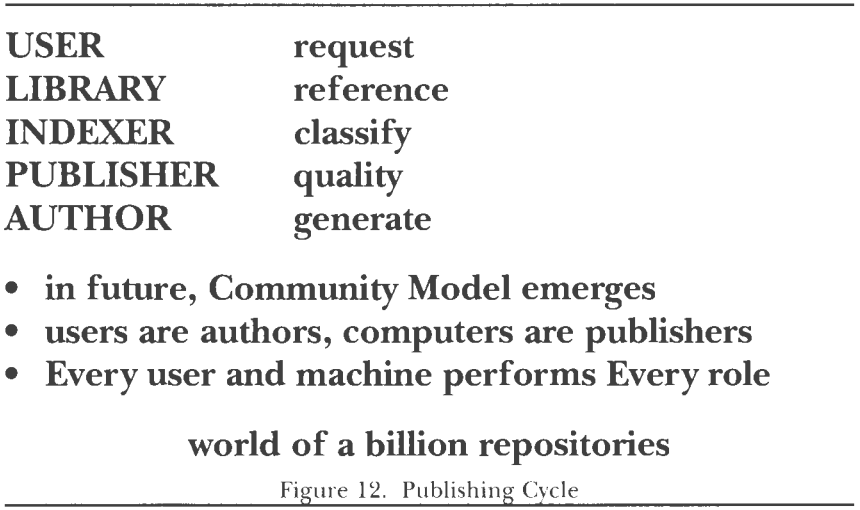


Figure 12. Publishing Cycle

That is only part of the story because the other strong technological trend is that the publishing cycle is breaking down. It used to be "here is the big library and here is an author, and the big library sits on a big machine, it is a big server, and the author sits on a small machine, it is a little client, and occasionally the author is going to shoot something over to the library." Then there are many people that are accessing this big client, so it is big things and little things, with the little things being users and the big things being libraries.

Well, that is not how the future will be. If you want a good illustration, let me just say that there were 100 million copies of Windows 95 in 1996, and there will be a publish command in Windows 95 that will basically, as part of the operating system, take whatever object you are working

on—like a spreadsheet or a word processing document—and place it on a Web site and index it. Thus every person will be able to easily publish things from their usual programs on the fly. Now it will not be refereed, it will not be a journal publication, but it will be somewhere. There will not be the current difficulties where you have to read a book that tells how to set up your publishing site.

So, this whole cycle that goes from users to librarians who do reference to indexers who carefully classify things to publishers who do the quality control for authors who generate the actual materials is going to break down completely. There will be individual computers and people who will do all those stages in different combinations, and the combinations will vary. But every person is going to do publishing, every machine will too, and every person will do some combination of the stages of the publishing cycle.

So, the Net of the future will have many levels of publications. You will have some personal documents. You will be the editor of a few small newsletters or clubs. You will be part of some professional societies, and each will have a professional letter or journal, because that is a big enough community that there will be sufficient numbers of people to be worth indexing in a more professional way. And so on to ever larger communities.

The end result will be a world where there are a billion repositories (and a billion might be a small figure). The ten-year projection is that there will be a billion personal computers, and each personal computer is going to have a couple of collections, so maybe 10 billion is a better figure, but a billion sounds like a big number. A billion is many more than the number of databases on Dialog. And a billion is a lot more than the number of sources you find in the index of all the databases, which is more like 10,000. This figure is a number that one can handle by just searching the descriptions in the index of databases. A billion is not a number like that. You need a *completely* different architecture to handle the world of a billion repositories.

This is not, I should emphasize, science fiction. This is straight technology extrapolation. There *will* be a billion repositories and, if the systems are ready and if the people who know about information retrieval do something, then maybe people will be able to find things in the world of a billion repositories. If they do not, then it will be, not like the Web now where you can actually find something if you are sufficiently energetic, it will be like you are in the Library of Congress, in all the archives that are underground that you know are unsorted. There is no card catalog—nothing. And you would like to find some information. What you are going to do is wander around at random and pass on the way some skeletons and occasionally hear somebody just before they die say, "oh, look over there in that catacomb and you might find something."

- **from Library model to Community model**
- **peer-peer not client-server information systems architecture**
- **support Net navigation and analysis**
- **cross-correlation from multiple sources**
- **design/implement Interspace environment**
- **part of Illinois DLI and CAN research projects**

Figure 13. New Architectures

That is what is going to happen. And the question is, Can technology solve the problem? And I am going to say, since I am a revolutionary technologist type, that the answer is yes. And I am going to tell you about some technology that might solve the problem. So, let me just emphasize that full-text search will not solve the problem and known semantic retrieval that works on 200 documents will not solve the problem.

We are now into the speculative revolutionary area of this discussion. What is needed are new architectures for systems that actually do something about analyzing and cross-correlating from multiple sources, because the library model where you have a few big things is totally blown away. What you have is a community model where there are a billion repositories, and they are all different sizes. For example, there may be a repository about cats. There is one about white cats; there is one about white cats with blue eyes that live in your neighborhood. And each repository is maintained by someone who is passionately interested in it. If you do not believe that there are such people, you have not used an electronic bulletin board or browsed the Web recently, or gone into a clubroom and looked at all the newsletters. That is what people do, so you have to deal with that world.

What will be described next is actually the backroom laboratory of the Digital Library project at Illinois and also the CAN which is a NASA information infrastructure project, and that is why it is being funded with high technology. But the funding agencies do not believe it is going to work.

Well, the easy thing is doing navigation and grouping, and that is what you can see the Web starting to do—i.e., within the Web, one can use a path to many different sources, and there are beginning to be facilities to record the path itself so you can play it back later. This is part of the facility required for what Vannevar Bush called "trailblazing" in his Memex Paper—if you are familiar with that—and what librarians call pathfinding.

*symmetric relationships*

- **navigation paths**
- **path recordings**
- **groupings: query sets and user lists**
- **selection and analysis**
- **path matching as basic retrieval**

Figure 14. Navigation and Grouping

For the full facility (well beyond what is currently available), one can edit the path so it says that you have been to different locations and the combined information is a valuable set. You also can do other groupings so you can do things that the telesophy system supports. For example, if you do a query, you can edit parts of it and save it as something you might wish to use later. One can make lists of interesting things that were found and were not even a path but just were gathered over time. And the reason to do that, first of all, is the convenience of having some way of recording things that were found in your searches because it is understood that regular indexing is not going to work. So, this is like recording reference sessions in order to re-use parts of previous works.

The second and most important reason for this type of searching is that paths are how a search *should* be done. While working with molecular biologists on WCS, after I had warmed them up over an appropriate number of years, they would tell me what they would really like. They would say, "I'd like to say I'm working on my own little organism, and here's three genes that are really important and here's the section of the map that I care about and here's some sequences in that section and here's three papers that are very important for this gene function—find me a similar collection of genes and literature that are in some totally different organism that is much easier to experiment on so I can do the experiment there, figure out what is important to do, and then go back to my more difficult but more interesting case."

You see that this is a general facility—i.e., path matching as the basic retrieval. What you did was search through the Net and hit some things that were interesting, and you want to find other groupings, other paths, that are similar. That is not full-text search. That is not even graph matching, although I basically described paths as graphs. It is some other kind of powerful semantic retrieval that nobody knows how to do.

There is a way of doing this powerful semantic retrieval, and that will be discussed next. The task is to handle repositories at a very fine grain level. So, when I say "repository" here and talk about organized collections, I do not mean what the Digital Library project is doing,

- **repository is an organized collection**
- **documents and indexing**
- **DLI establishing large repositories for major publishers**
- **WCS integrated large formal (journals) and small informal (bulletins)**

- **need semantic retrieval for across publishers**
- **need semantic classify for small publishers**

Figure 15. Community Repositories

which is making a collection for the IEEE journals. There are trained professionals who do that, and I do not really mean what WCS did—what the Worm system did—which are things like specialty journals and things like the community newsletter for several hundred people—e.g., you and your neighbor with the cat who have a newsletter about the cats in the neighborhood.

You and a small group form a community—e.g., I take my daughter to a music class on Saturday morning that has five other kids who are two years old and five other parents. That set of people has a common interest—i.e., they would like to have a collection of their information on kid-related topics that they could search, and I would wager that there are similar sets of five parents elsewhere who would also like to be able to access this collection. As one of those five parents, I am willing to spend a modest amount of time making a collection and doing some indexing, but I am not going to become a professional indexer like Inspec would hire in order to do electrical engineering.

So, what you need is some way of really being able to do classification for small publishers and some way to use that classification to search across the collections at a deeper level. The collections will span numerous and different publishers, from really little to really big and from really low quality to really high quality. It is difficult to tell how the quality level might vary. The small ones might actually be more carefully done than the big ones, but the professionalism is different. So, how are you going to be able to search across all those collections?

First an examination of what professional classifiers do now. Those are human indexers. They make a subject classification of the important terms in an area and indicate which terms are bigger and littler—this subject hierarchy is correct in some profound sense. The hierarchy represents the meaning of the subject area. However, the terms tend to be very general. For example, in the Worm Community System, we obtained

- **human indexer (manual classification)**
  - **—hierarchical terms (classify) correct but general**
- **machine indexer (automatic classification)**
  - **—related terms (co-occur) specific but incorrect**

- **manual yields meaning for precision**
  - **—interactive interface for Inspec thesaurus**
- **automatic yields context for recall**
  - **—co-occurrence matrix for 400K Inspec abstracts**

***multiple displays for different classifications***

Figure 16. Indexing and Classification

a copy of *MeSH*, which is a very well done thesaurus covering all biomedical research generated by the National Library of Medicine. We were all excited about it until it became clear that every article in the Worm literature—all 5,000 of them—had exactly the same *MeSH* terms.

That was actually what started us down the path toward automatic indexing. It made us think that even for this collection—i.e., a couple of hundred people so that it is a reasonable size community and not a couple of neighbors—that you needed better technology. So we started looking at co-occurrence matrixes that record the frequency of terms occurring together. This statistical technique goes down to the real words in the documents and can be done automatically but has nothing to do with meanings. It is a context of some kind, so it is relatively good at recalling things but not so good at being precise—i.e., the automatic technique is quite specific but not necessarily correct.

Following up on this work from WCS, as part of the Illinois Digital Library project, we built an interactive interface to the indexes from both manual and automatic classification in electrical engineering. The manual classification was the real Inspec Thesaurus—10,000 terms carefully done by professional indexers. You can use a graphical interface to this classification scheme to move up and down the subject hierarchy and then find desired words and use them for search terms. That is very helpful to see what the main categories are, but it is not very helpful for discovering the actual words appearing in recent papers because they just are not in the thesaurus.

So, as before, we also generated an automatic classification scheme by gathering statistics of which terms occur together and how frequently. The interactive interface to this "concept space" suggests alternative terms

for which to search—i.e., given a word, it gives a list of other words that occur most commonly with that word in context. The context words are all intermingled—bigger, littler, useless, useful.

This co-occurrence list is not meaning—a professional indexer would reject this completely (and they have when we talked to them)—but the context lists are practically useful as search suggestors. This is, in part, because the granularity is much finer—there are 100,000 terms from the same Inspec corpus (ten times more), so that you get not just "deductive databases" but also "Prolog" and "inference mechanisms," and partially because the system is interactive so that the users are perfectly happy to sort through the lists themselves deciding what is useful in exchange for getting the full range of potentially related words from the documents.

What we found in molecular biology—in small experiments in molecular biology—is that the concept spaces are pretty good as memory joggers. The fact that you can also generate them automatically is really convenient because it means you can use these in cases that are inappropriate for professional indexers. I'm thinking about the cat example. You can get professional indexers to work on a repository for journal articles in electrical engineering but not on a repository for notes on the cats in your neighborhood.

- **automatic indexing of concepts**
  - **—find context of phrases within documents**
  - **—generates a concept space based on term frequency**
- **useful for interactive searching**
  - **—given a term, can suggest other terms**
  - **—merging concept spaces supports vocabulary switching**
- **concepts require supercomputing**
  - **—concepts space for Inspec took 1 day on SGI Challenge**
  - **—co-occurrence matrix for 400K abstracts**

Figure 17. Semantic Retrieval

Now some further discussion about the automatic classification scheme since, if the manual one is there, you should definitely use it. The automatic classification techniques all are statistical correlations of the context within documents. The particular one we are using is co-occurrence matrixes, which is only one of the one hundred ideas regarding how to do deeper semantic retrieval that have been in the information science literature since the 1960s (when I say "we" here, I mean my colleague Hsinchun Chen from the University of Arizona and myself).

But co-occurrence is one that is now computationally feasible if you have a supercomputer. For example, if you take the SGI Power Challenge, a high-end supercomputer at NCSA, and take a day of computer time, actually twenty-four hours, then you can compute a co-occurrence matrix of a real collection of 400,000 abstracts. That was not true in the 1960s, and it has nothing to do with the algorithm being better, although it is tuned a little bit. It has to do with the fact that computers are enormously faster and so some of these old deeper semantic techniques can actually do something real. Since techniques—like co-occurrence lists—are useful as term suggestors, this might be a real break into semantic retrieval.

There is no magic, no natural language parsing, no fragile domain rules. We have a lot of computational power, and we can look at the word frequencies ad nauseam. This is just a first attempt to develop deeper semantic retrieval. So, you get terms like "Horn Clauses," which is a really fine-grain technical term in deductive databases, and you get terms like names of people that write articles about deductive databases. In molecular biology, you get names of genes that occur commonly in articles about that particular concept, which is very helpful to users, especially since an indexer would never put the name of a gene in the *MeSH* thesaurus. So, much of the power of this particular technique is that it does not have any semantics in it at all—it just takes whatever words are there. You hope that there is some guilt by association, and that if two terms occur in the same context frequently then one is a good alternative for the other when doing a search.

Now another nice thing we did, we tried computing the co-occurrence matrix of several different collections because the Worm system was actually several different collections, in different orders—just because we were curious. We had done it in one order and then we said, "does it make a difference if you do it in the other order?" thinking it does not make any difference. And the answer is, the lists were completely different. Then the first thing that occurred to me is that maybe the vocabulary problem could be solved.

Now, that is not to say that the vocabulary problem is solved, but it should be explained that this is a small development against that. The vocabulary problem is that you have the same concept in two different subject areas but the terms are different. So, in engineering for example, "fluid dynamics" is a term that occurs in many subject areas, but the words are completely different even though the concepts are similar. Could there be a system where you say: "I'm a civil engineer who designs bridges. I'm interested in fluid dynamics to compute the structural effects of wind currents on long structures. I think ocean engineers who design undersea cables do similar computations for the structural effects of water currents on long structures. I want [the system] to change my terms for

- **fine-grained concept spaces**
  **—for every community and subcommunity**
- **user and collection modeling**
  **—choose domains for user and for search**
- **interactive vocabulary switching**
  **—intersect at common terms to suggest across domains**
- **supercomputers as time machines**
  **—personal computers same computations in 5-10 years**

Figure 18. Vocabulary Switching

talking about fluid dynamics into the ocean engineering terms and search the undersea cable literature as automatically as possible."

Well, that is basically what vocabulary switching technique does. It allows one to make this fine-grain concept space, built on co-occurrence matrixes, for a very small collection, so you can do it for really small communities. You can then, on the user end, say "I'm in these three communities, that's what I know about, and I want to search these other three." Then the system will automatically intersect the corresponding matrixes, which are just concept graphs, and let the user interactively switch the vocabulary from one space to another to facilitate the searching of the desired community repositories.

It is possible to do these computations now with supercomputers. You should know, if you are not accustomed to them, that the significance of supercomputers is that they are good as time machines. It is well known from the technology curves of the past, which are probably too slow, that whatever speed a supercomputer runs now is what a $3,000 desktop machine will run in ten years. So the following experiments are a conservative estimate of what you will be able to do in ten years.

So, here are two vocabulary switching experiments. This first one is actually going to appear in *JASIS* very shortly, and it did two areas of molecular biology: worms and flies. It had about 5,000 documents in each area and each took about ten hours of computation on a workstation. So, you could say, for example, "here's sperm about worms and sperm about flies" and it would list twenty-five terms. Ten of these would be the same, so you would ignore those and look at the different ones.

Vocabulary switching is needed for this sperm example. It is known that worms have odd sperm—I happened on the Worm Community System to work with the world's expert on worm sperm. Worm sperm has little pseudopods and crawl like amoeba, while all other sperm in flies and

- **small-scale in molecular biology (JASIS)**
  - **—worms and flies**
  - **—5000 documents generate each space**
  - **—10 hours per space on a workstation**
  - **—try "sperm" as connection term**

- **Large-scale in engineering (in progress)**
  - **—5M abstracts from Compendex**
  - **—1000 spaces across all of engineering (5K per space)**
  - **—1/4 hour per space on a supercomputer**
  - **—try "fluid dynamics" as connection term**
  - **—fine-grain subdivide for user-driven switching**

Figure 19. Switching Experiments

everything else swim since they have these little flagella that wiggle. So, what you want to do is change all the crawling to swimming and change all the pseudopods to flagella.

Well, if you look at these co-occurrence lists (for worms and for flies) and you look at the different terms, then sure enough, crawling and pseudopod are on this end and swimming and flagella are on that end—this is not automatic. You have to realize that because in with those two good terms are ten ridiculous ones that are way too general, and ten of them were common, you ignore those—i.e., of the top twenty-five terms for sperm in worms and flies, about ten are common to both lists—about ten are useless, leaving only five as potentially useful.

Returning to the original example in engineering, what this technique would do if it was fully automatic is it would say, "Oh, no problem, here's the three terms you really care about. They're these three relevant terms to you in undersea cables." That is not what it is able to do right now. What it is able to do right now is say, "Okay, here's the terms you want to search, and here's what you know about, and here's what you'd like to know about. Here's ten terms that you know, and here's ten terms that you don't know. Match them up yourself."

And that is amazingly a great deal of help compared with nothing. With biologists, they are really fast at scanning through lists of potential terms and are very grateful to have a list of possibilities, because otherwise they are going to sit there and try words at random. The concept of space intersection is far from perfect or even correct, but it is much better than trying words at random. If you have ever been in a reference library, you know how bad people are at actually searching.

So, encouraged by that, we tried a real experiment. We took 5 mil-

lion abstracts from Compendex, which is an index covering all domains of engineering, and generated 1,000 spaces of roughly 5,000 abstracts each. Thus each space approximates a community repository of the same scale as the ones from actual communities in molecular biology, and the intersection of the spaces simulates an interspace across all of engineering. What we are actually going to do is divide Compendex by class codes so that the size of a space is a fairly fine-grain subject domain like bridges or highways.

A simulation this large can only be run on a supercomputer. Even so, at about 1/4 hour per space on a machine like the Convex Exemplar, plus intersecting the spaces, this is still going to take about two weeks of computer time. Fortunately, the newest and largest supercomputer that NCSA obtained is still in its testing phase, and I was able to persuade them that this was an interesting application, so we are able to reserve the time to try this as a hero experiment.

Then the question is, can you issue a query like "fluid dynamics" and really do useful interactive vocabulary switching? That is totally unproven, but it works much better than you would expect in molecular biology, where it really does do something. It is computationally feasible and it does something, and that is much more than not being computationally feasible and not doing anything.

- **domain experts but classification amateurs**
  - **—large community indexing is too general and too old**
  - **—small community indexing is not consistent**
- **useful for interactive subject classification**
  - **—automatic suggestions for potential classifications**
  - **—domain expert culls list from "controlled" vocabulary**
- **semi-automatic support via concept spaces**
  - **—concept dictionary of tag words from co-occurrence**
  - **—tag frequency in documents determines classification**

Figure 20. Computer-Assisted Indexing

The other side of all this is, can you also use concept spaces to help with indexing? Well, what is the problem with indexing? If you can get a professional indexer, they will do a good job for their broad subject area. That does solve the problem of electrical engineering, but it does not solve the problem of bridge swaying or neighborhood cats, which are too small a subject area to afford a professional indexer. And for these specialized repositories, the terminology in the professional subject indexes is too old and too general.

However, if you try to solve the indexing problem for specialized communities by letting individual people from that community do indexing, you discover the value of using trained professionals. As many experiments have shown, ordinary people have really wide variations in how they classify things. An ordinary person will not even assign the same terms to the same document twice, much less will two different people assign the same term to similar documents, which is what you want. Remember, all we can do is string matching underneath, so the indexing has to be precise and consistent. There is no magic here.

So, suppose you could have an automatic program that would suggest topics for classifying a document and then let a person correct the list? For example: "Here's twenty-five terms that this document should be about. Choose five from that list." The domain expert, who knows about bridges or worms or cats, can do that. They know the subject area and the meanings of the terms, so if the system could suggest consistent terms to limit the variation, you would get an interactive indexing system that enables amateurs to approximate the quality of professionals. This ought to sound similar to the sort of solution that concept spaces provide for semantic retrieval.

We are currently trying a set of experiments that basically provide a domain-independent version of the old 1980s technology that used to look through newspaper articles for the CIA and try to identify which ones are about revolutions. What these old systems did is use tag words. So they said: "Revolution has these ten words that commonly mean revolution and tank has these ten words and spaceflight has these ten words. This document mentions three words for revolution and one for spaceflight so it's about revolution." As you might imagine, they would get fooled quite a bit, but they would often be able to assign what topics documents were on and some of these were right while some were wrong. So they would say: "This article is about tanks and spaceflight," when it was about the Russian invasion of Hungary because it mentioned the word "satellite" a lot and satellite was a tag word for spaceflight.

This concept identification technique relies on having a concept dictionary giving the tag words. Well, the concept "space" really has that. It says: "Deductive databases—here are ten words that might be useful, that commonly occur with deductive databases, so if you see one of these, the document is probably about deductive databases." If you use the concept "space" as a concept dictionary and look at the words that commonly occur together as tag words, then you are able to make a suggestion list of which words could be used to classify the document, just like a professional indexer will choose some terms from a controlled vocabulary like Inspec or *MeSH*.

It is unproven what will happen with this. The experiments are just starting, but the basic idea is sound in the sense that it is an automatically

generated controlled vocabulary specific to a particular topic, and the actual selection is done by a subject matter expert.

So again, like all the things in this discussion, in the future part, this is something that is sensible, that might actually work, and even if you do not believe this one, it may be that some variation on this will allow the ability to do indexing. Remember, if you do not do fine-grain indexing, you will be unable to find anything in the world of a billion repositories.

- **every machine has its own information space**
- **every machine has its own concept space**
- **spaces for every user and every community**
- **search is matching selected objects**
- **relies on computer-assisted indexing**
- **analysis is merging community spaces**
- **vocabulary switch through graph intersect**

Figure 21. Applications Environment

So, now it is time to discuss Interspace software very briefly. It is an applications environment built on top of the Internet, assuming that the Internet has evolved into a worldwide object-oriented operating system. Basically, it assumes that every community has an information space, every information space has a corresponding concept space, and then the Interspace is the intersection of all these spaces.

The environment for the Interspace supports searching and analysis. The searching is what I said before. You select a group of objects and the environment locates similar groups. Vocabulary switching is done automatically. So, this is just a whole network information system that uses this vocabulary switching and concept spaces to try to do semantics at a fine-grain level in order to handle community repositories.

- **objects** —**fine-grain manipulation**
- **navigation & grouping** —**path recording**
- **retrieval & classification** —**concept spaces**
- **correlations** —**path matching via concept spaces**

- **prototype in Smalltalk, CORBA, ObjectStore**
- **application in digital libraries, GIS**

Figure 22. Interspace Prototype

This is actually what my lab at the University of Illinois is doing—building a prototype of the Interspace. Kevin Powell and I wrote an architecture document laying out all the parts of the environment, and my team is in the process of implementing the first full prototype.

If you want a little technical detail: it has objects, it does retrieval, it tries to do correlations. The prototype assumes a distributed network of objects by using high-end software technology like Smalltalk and CORBA and is constructing an applications environment to handle the concept spaces and semantic retrieval. We are then going to try some sort of hard applications where there is much data and easy questions have hard answers which require looking through lots of things to cross-correlate like digital libraries or geographical information systems. Over the next few years, we will be evolving the software and simulating the world of the Interspace, with spaces like the thousand community repositories in engineering discussed earlier.

- **Beyond Search to Analysis**
- **Cross-Correlating Information from many sources across the Net**
- **The Net solves Problems**
- **Every community has its own special library**
- **Every community and every person does indexing!!**

Figure 23. The 21st Century: Analysis

To summarize, the claim is that the twenty-first century is going to go beyond search and into analysis. And what analysis really is is cross-correlating information from many sources. And then what you will be able to do is solve problems, not just find things at random. And in order to do this, what you need underneath is very fine-grain classification. That is the only known way of handling the world of a billion repositories.

What that means is that every community, large or small, has its own little digital library. The software does some computer-assisted indexing, and it has some "semantic" retrieval that uses that indexing to try to do vocabulary switching, to try to do better kinds of search, so that there is more responsibility for some individuals to develop collections, but this also means that the average person might end up being sort of a librarian. They might maintain a collection. They might do searches on an everyday basis.

So, what you need is to embed some of this higher-end technology into the standard network software that ordinary people use in order to be able to do this new kind of functionality. This new functionality will happen. Commercial pressures will force this to happen. The real ques-

tion in the world at large is: "Is the Interspace going to empower the individual person so that they will be able to actually find things and solve their own problems and maintain their own collections, or is it going to be yet another new medium for providing more advertising to enrich the greedy evil corporations?"

***everything goes in with transparent manipulation***
***everyone gets credit with community sharing***

**merging spaces from other communities:**
- **molecular biology**
  - **(coli, yeast, worms, flies, mice, men)**
- **neuro biology**
  - **(moths, mollusks, rats, cats, monkey, man)**
- **other sciences...**
- **other domains...**

Figure 24. Building the Interspace

So, suppose this all works? Suppose it is ten years from now, and everyone has something that supports the Interspace technology on their desks and in their homes. A box that comes with the software environment built in and a plug into the Interspace. Just like a set-top box comes now with Netscape and a cable modem. Well, what that means is everything in the world goes into this space: everybody can share to put things in, everybody can browse to get things out. Then what really begins is building the Interspace—i.e, creating all the individual community repositories, connecting all these individual spaces together.

The most likely start will be in science and engineering, because those people are comparatively rich and they are the ones who have the high-end technology. Just like the ARPAnet begat the Internet and government-funded labs begat the Web, the same stages will happen with the Interspace. That is why I provided examples from high-end digital library research.

The next thing to happen is to start merging individual community spaces such as those I have discussed in biology. For example, start with molecular biology (worms and flies to mice and men) and on to neurobiology (rats and cats to monkey and man) and on to other sciences and other subject domains.

Well, there are regrets that I am not doing the Worm project anymore, so I cannot say "today the Worm, tomorrow the World," which is how I used to end talks. But what I have to say is that there *will* be a WorldNet whether one likes it or not. Every community, from really big

**every community repository,**
**large and small**

**living in the Interspace of**
**all the world's knowledge**

Figure 25. Building the WorldNet

ones to really small ones, will have a nice collection. It will be indexed. There will be ways of accessing it and correlating it. So what you will begin to see is that there really will be an Interspace.

This will be where people live. You see things now that say people live on the Net. Well, that is true of a few specialized people who are questionably human beings, a group of which I am a member, I am sorry to say. But this will be true for the average person. Just like television became ubiquitous, the Net is the world of ten years from now. So, you have to get ready for it and figure out what you can do to contribute to it—to make it help people by letting them get the information they need to solve their problems and being able to organize their own collections rather than hurting people in ways that can be easily imagined.

The NII is often referred to as the best technique for selling advertising for 500 channels of mud wrestling. Maybe now with the Web it has become the medium for selling advertising to access a million home pages of dogs barking. That is not, from a purely personal standpoint, the appropriate use for such a far-reaching new medium. The vision of the pioneers was always education, not entertainment—the Net should become the way that ordinary people solve their problems. This new research technology might be the way toward that vision, toward the Interspace.