National
Intelligence
Council

DRAFT

Open Source Task Force: A Vision for the Future

Background information for Task Force meeting on 14 January 1992

Comments may be directed to the NIO/STP on secure 37296 or Stu 703-448-0616. (U)

January 13 1992

Important Preface

This draft task force report contains a vision for an open source program. This vision is substantially different from the current program. In essence, it is of a program that gives Community analysts access to far more open source data, thereby making them far better informed and thus both more efficient and more effective in their jobs.

The vision presented is well founded on social, as well as technical developments proven outside the Intelligence Community. But because many inside the Intelligence Community are not familiar with these developments, feasibility is challenged at every level. "You can't store all that material." "You can't communicate all that material." "You can't search all that material." "It's too expensive." When those are addressed, the next level of criticism is "You can't afford to get the data into the system." Finally, the more social issues: "You will need a computer guru to use it," or "No one will have time to use it."

It is not the intention of this report to specify an architecture for an open source system. But proof that a vision is possible, and that the above concerns can be addressed, requires something more concrete than a simple vision statement.[1] Therefore, this draft report leans closer to a system description than just a vision statement. However, the intention is not to limit the vision or an architecture based on a vision—it is merely to lend some credibility to the reality of a vision.

[1] The converse is that defenders of the status quo feel comfortable that proof of their concept is easily demonstrated.

Key Judgments

- **Intelligence for the 90s and beyond will require far greater access to open source materials than today.**

- **The technology exists to support an open source program many times larger than today's at a reasonable cost.**

- **A modern open source system can greatly enhance the utility of the data by allowing users to share views and annotate source documents.**

- **Plans for the program could be drawn up quickly, with initial installations within a year.**

- **Recommended actions include measures to:**

Open Source Task Force Report

The Problem

The world is changing. The Intelligence Community's information management systems were designed to function most efficiently against the closed Soviet Union. Today, the needs for intelligence are much more worldwide. As examples, the worldwide proliferation of weapons of mass destruction and economic competitiveness analysis require different kinds of data and in vastly greater quantities than most previous intelligence topics.

Along with the political changes, there have been many technical advances that allow us to think of whole new ways of doing our business. Much of our current open source architecture was developed in an age when information processing and communications were in their infancy. This led to the need, for example, to make lists of key words for documents for easy retrieval and to abstract documents to save on storage and communications costs.

Add to this the advances in publishing and commercial data bases, which have led to an increasing percentage of printed matter available somewhere in the process in electrical form, and we have the makings for a revolution in open source processing.

Making a revolution happen, though, requires knowledge of where we want to go. Thus, a vision statement. But a complex project requires vision at several levels. Two are presented here.

Vision Statement—Management's Viewpoint

The Intelligence Community needs an open source information gathering and data management program that assures that judgments in intelligence products are informed by all relevant open source materials.

Vision Statement—Analysts' Viewpoint

The good news is that developments in the commercial sector give us a framework for a vision of a robust open source program for the Intelligence Community. As viewed by a user, such a system would have the following attributes:

— Upon logging on to the terminal on his desk, the user would have immediate access to a broad range of information services, including:

- Worldwide wireservices—regardless of language
- Worldwide newspapers—regardless of language
- Radio and television reporting
- Books and periodicals
- Electronic databases, etc.

— But while this access is important for research, to verify reports, and to check hypotheses, it is not sufficient. The large volume of this information requires that the analyst has help in searching and making correlations. A number of such aids are at his disposal:

- An organization of the data sources that allows the analyst to look only where he expects results.
- Smart profilers that recognize word associations as new data comes in.
- "Expert intermediaries," who specialize in following particular data sources for highlights.
- Colleagues around the Community who come across relevant data and report it to him, along with comments on its significance.
- An electronic bulletin board where he can post queries to colleagues.
- High speed tools to rapidly search data source subsets.
- Software "agents" (software lifeforms that are to a computer virus what a human is to a real virus) that the analyst creates to search far and wide, 24 hours a day, for significant correlations in the data sources.

— Also, because open source is but one class of data used in analysis, the access, correlations, etc., should be "seamless" with that from other data sources and easily incorporated into the publication process.

Where We Are Today

The current open source architecture has served us well. However, when open source is viewed in its totality—from collection to use—there are significant shortfalls for this new world:

- The quantity of data provided to the Community's analysts barely scratches the surface of what is needed.
- Open source materials that are collected are often not made available to the user in a way useful to the user.
- Collection is highly tasking specific, making it good for addressing the specific question, but not good for long term research, competitive or retrospective analysis, or for institutional memory.
- There is no way to incorporate "value added" from users to other users.
- Users have no way of knowing what open source data is available to them or where even to look for much of it.
- Users have a poor ability to search large open source holdings in a timely manner.
- The Community has no standards or capability to deliver foreign language materials to users in electrical form.
- There is no way to get printed material scanned and entered into the electronic holdings.
- Current electronic holdings are limited to text only. There are no standards or capabilities to store and make available to the user images, graphs, or voice materials.

Can the Vision Be Implemented in Our Lifetime?

Perhaps surprisingly, much of the architectures and technologies to implement the vision are available today. Conceptually, the architectures of the CompuServe and Internet systems can serve as models, as will be shown shortly.

It is possible to see in operation today at least the beginnings of all the features

in our vision. Some features, such as widespread collection, storage, and retrieval of television will have to await the next generation of storage and communication technologies—about ten years from now. But the basic architecture and much of the functionality can be built now, at a reasonable cost and at low risk.

What Would It Take to Implement the Vision?

The last ten years have seen a revolution in the technical tools for handling large amounts of data. In the past, we might have considered a central repository (computer center) for collecting and archiving open source materials. But, with the availability of powerful desktop computers, we have learned that limitations inherent in central systems can be avoided by going to massively parallel systems. That is, for some classes of problems (which includes the problem of an open source data system serving many users), the overall power of the system can be effectively multiplied by simply increasing the number of computing elements incorporated in the overall system.

The parallel or distributed system approach has several important additional advantages. The overall system can be upgraded gradually—simply by adding to or upgrading the individual elements—as technology advances. Moreover, the communications problems are minimized because the users' terminals will be communicating with numerous other computers, not all having to go to one central computer. Also, a distributed system is more tolerant of failures.

The conceptual system about to be described is viewed from the users' viewpoint. A lot of capability will be "assumed," in the system. But following the description of the system as seen by the user, we will look at the driving technologies (hardware and software), and compare them to where we stand today.

A Users' System

When approached with the thought of having access to much more data, most analysts in the Intelligence Community envision logging on to the new system and seeing a report that his profile has identified 3,562 new cables for him to review. No thanks! Isn't there someone (for example at FBIS) who can serve as an "expert intermediary?"

Any new system bringing vastly larger quantities of data to the user must account for the fact that most users won't have a lot of time to spend on the terminal, and may not be very sophisticated in computer operations. By the same token, however, the system should provide the sophisticated user, whose job requires it, the ability to take full advantage of the resource.

When the user logs on to this new, conceptual open source system, he sees a screen like that in Figure 1.

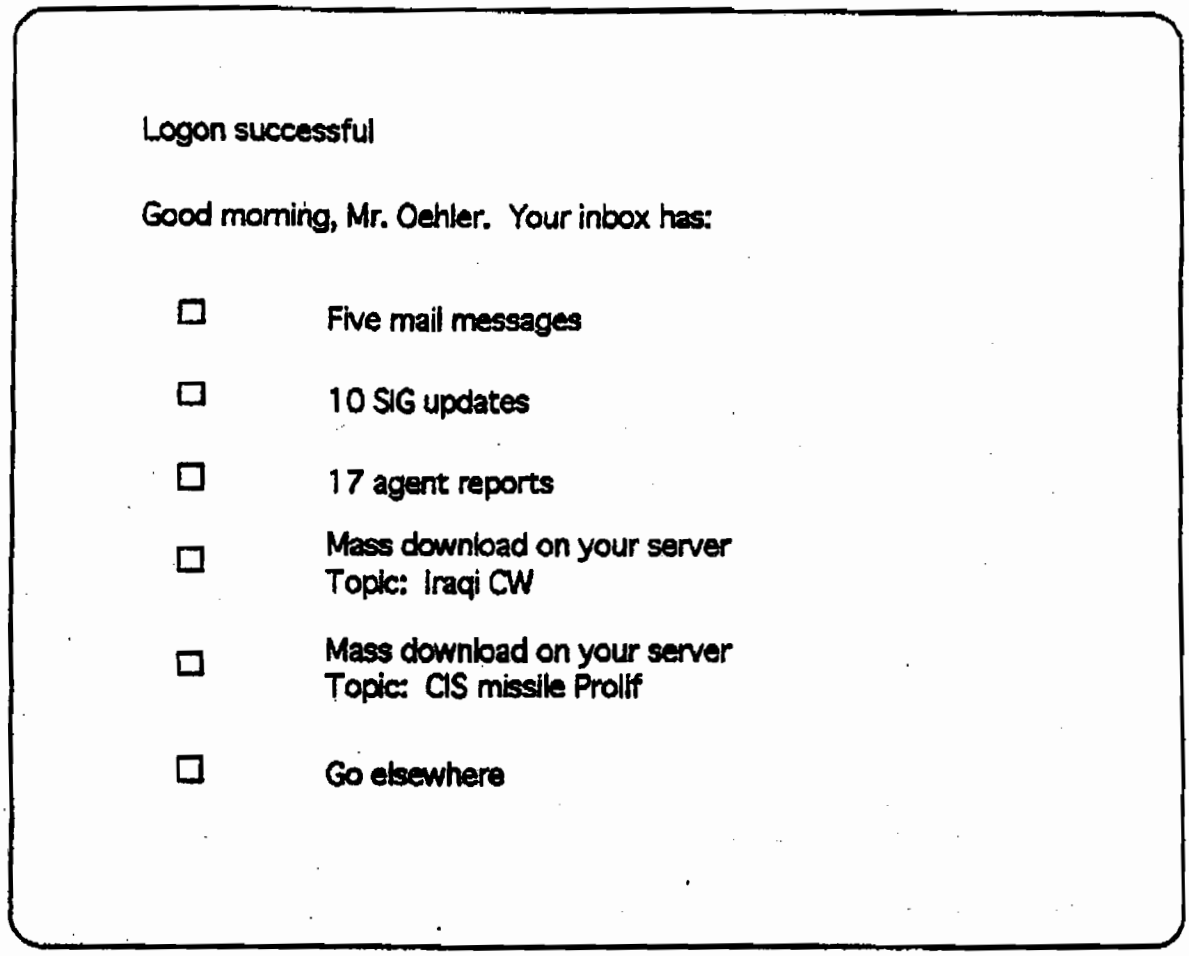


Figure 1

The options that he has available are:

Mail messages — informs him that he has five new electronic mail messages. He accesses these by clicking on the box to the left.

SIG updates — is the heart of the open source system. SIG stands for Special Interest Group—to be described shortly. Clicking on its box will allow him to access the SIGs he has chosen to monitor.

Agent reports — tells him that the "agents" he has created in software to forever roam the data files have discovered something and want to report back. This is the only part that is a little futuristic, but both CompuServe and Internet have a start on this. This will not be discussed further in this report—it isn't needed to show the power of the system.

Mass downloads tells him that the results of his request for a lot of data on a broad topic has been fulfilled, and the results are stored on his local server computer for his perusal. This feature allows him to have the system gather material needed for analysis. For example, the user might have requested all the material from many different files that contain the words "Iraq or Iraqi" and "CW or chemical," from 1982 to 1990. Such a broad search could result in tens of thousands of pages of data downloaded to his local server. Now, when he wants to, he can use his local text search tools, and nose around the data with near-instantaneous response time needed to do his research.

The reason for this feature is that it gives the user access to a very large data set to search, without requiring the entire open source network to be able to respond to his needs with near-instantaneous access. The disadvantage is that the analyst must know ahead of time that tomorrow he will want to research a particular broad topic. This feature is available today on many systems.

Go elsewhere allows the user to go somewhere else—an area or SIG that he normally doesn't follow.

So far, simple enough.

As noted, the Special Interest Groups are the heart of the system. They are basically electronic bulletin boards, where messages are posted for any interested user to see. SIGs have the following properties:

- They are run by a System Operator (SYS OP). [The terms SIG and SYS OP are the terms used in the Internet system. They are perhaps not the best terms—a SYS OP, for example, is an expert in the substance of the SIG, not a technical computer operator. In the CompuServe system, the SIGs are called Forums.]

- They are basically postings to a bulletin board.

- The SYS OP decides what gets posted on the board.

- Some SIGs require a lot attention by the SYS OP, others little.

- SIGs are easy to create and retire. Any user on the system can create a SIG and be its SYS OP.

The Internet system has literally thousands of SIGs, covering all manner of

subjects. A system for the Intelligence Community might have a few hundred.

For our conceptual system for the Intelligence Community, example SIGs might include:

CIS Political Developments	TASS
North Korean Nuclear	French Wireservices
Iranian S&T	Reuters
Environmental Issues	FBIS Middle East Daily Report
GATT	FBIS CIS Daily Report
BW/CW Terrorism	US TV Transcripts
Japanese R&D	Korean Times
Middle East Oil Developments	Le Monde
FBIS Proliferation Watch	Guide to SIG Use
JTEC	Agent Development

While all SIGs are really the same in structure, there are differences in how they are viewed by the user. For example, the Reuters SIG would be a simple file of all the Reuters news stories. It would have very little interaction by the users other than reading the postings. The user might occasionally send a message to the Reuters SYS OP asking, for example, why yesterday's stories weren't posted—whereupon he might get a reply from the SYS OP saying that there was a technical problem, but that it would be loaded tonight.

A SIG like "CIS Political Developments" is very different. The SYS OP in this case would be a CIS specialist, who selects articles from other SIGs and "donations" from other users that he, the SYS OP, deems appropriate for his readership. In this case, the SYS OP could be the FBIS "expert intermediary" or could be an analyst in OSEA or DIA.

Is the SYS OP a collector or a user? The answer is that he is some of each—this blurring of the distinction between collector and user is important to the sociality of the system. There isn't a requirement that the SYS OP for this type of SIG be in this group or that group—just that he is a person interested in the subject and can manage the board. He may even have a number of assistants helping him keep

the board informative to other users.

The user interacts with this kind of board differently than with the Reuters board. The user, after reviewing the postings on the board, may have a question to ask the "expert" SYS OP, so he sends him his query. The SYS OP may know the answer directly, refer him to a posting made some time ago, or post a message on the SIG board itself to see if any other readers can help him out. If a answer comes in from another reader it, too, becomes part of the database, the corporate memory.

This interaction of the users in the SIG means that the board takes on a "value added." This "value added" may come from the analysts knowledge of how this information compares with that from, say, classified sources, and makes the board far more valuable as an analytical tool than just the data itself.

The ability to easily create and retire SIGs gives this system a very un-government-like feature. If the SYS OP of a SIG isn't providing a service to his readers it will be obvious because no one will be subscribing to the service. Someone else may start one on the same or similar subject. In other words, the marketplace will ensure that the good drives out the bad.

There are other kinds of SIGs as well. For example, a SIG on how to use SIGs, a SIG containing a directory of services etc.

Where are the SIG's located? They can be located anywhere. To the user, it is all "transparent," meaning he doesn't need to know. It doesn't make any difference whether a particular SIG is housed in the office next door or halfway around the world.

This transparency means that the SIG can be located wherever the best person to host the SIG is. This helps with the data storage and communications circuits because they don't have to be all in one giant computer room.

SIGs can be set up to address short-interest-time issues. For example, if there is a coup in a particular country, a new SIG could be created and all interested users can become "wired" together to exchange news reports or to coordinate intelligence production.

With that as background, the power of the system becomes apparent. The unsophisticated user would probably "sign up" only, say 1 or 2 SIGs; for example, the electronic version of the FBIS daily report for his area and maybe one interactive SIG. He may even have the material printed because he prefers reading paper. He expects that if there is anything significant going on in his area, someone will post it for him (the "expert intermediary" or a subordinate). The power user has access to all the data in the system.

The system is capable of handling foreign language materials, some images, and some voice files. An English-speaking-only analyst has the tools to search some foreign language materials for articles of interest and, if necessary, send them off to

another SIG for translation.

The SIG SYS OP is responsible for ensuring that the data in the SIG is archived for all time.

Conceptual Physical Architecture

A hardware architecture would could look conceptually something like that in Figure 2.

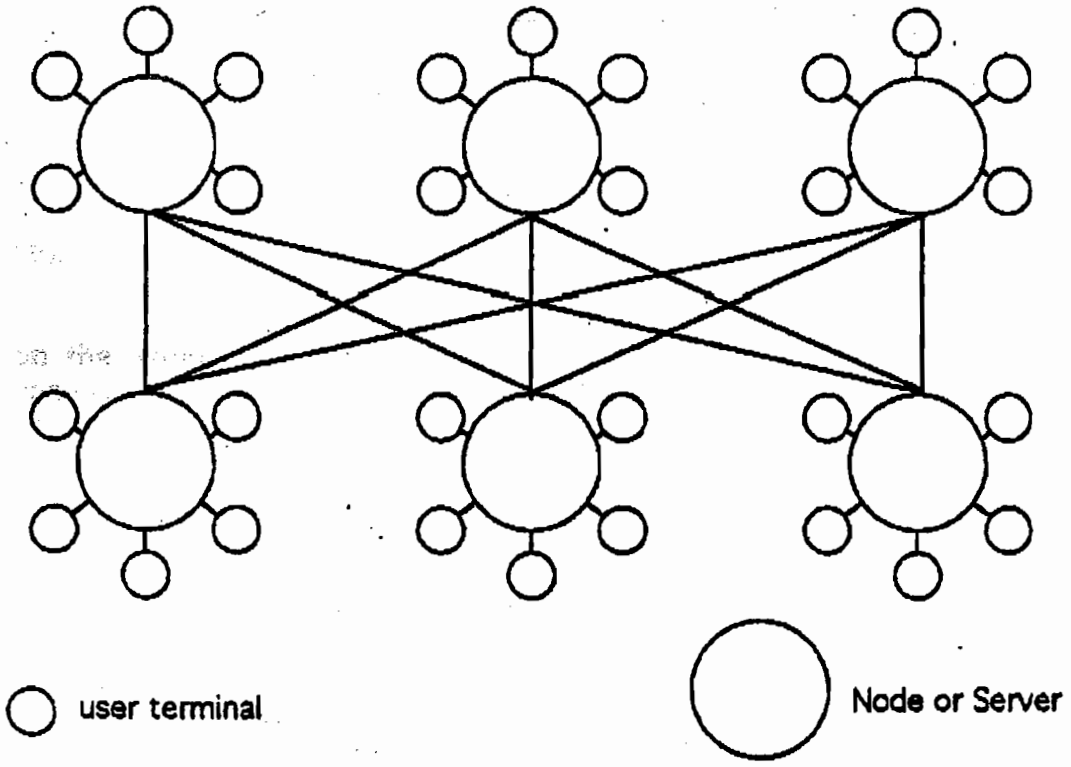


Figure 2

Each node or server is connected to a number of users through local communications paths. The nodes are all connected to each other, usually over long-haul, commercial telephone circuits.

A small group of users, such as State/INR might have only one node connected to a few hundred users. A larger user might have multiple nodes. All transparent to the user.

The data would flow as follows: If a user wants to receive postings on a

particular bulletin board, such as Reuters or CIS Political Developments, he needs to send an electronic message to the SIG SYS OP. The SYS OP then puts the user's electronic address on his mailing list. When a new message is posted by the SYS OP, a copy is placed on the user's server. When the user logs on, he receives a message from his local server saying that a new posting is there. If multiple users on his server were interested in the same SIG, just one posting would be made to the server, but all would be told of its existence.

Because the information is on his local server—not some far-off place—the response time to his queries is rapid, and he can use his own text retrieval and search tools. The response time demanded from the long-haul circuits is important only for messages to and from SYS OPs, on-line analysts' meetings, etc.

What Are the Technology Drivers? How Do We Fare Today? The Near Future?

System performance depends on a number of factors: The capability of the local servers to store the local users' data and to communicate with the local users; the capability of the users' personal desk top computers to perform his local search duties, the capability of the long-haul communications lines; the software connecting it; and not insignificantly, the cost of it all. Also, once the system is put together, can we afford to feed it?

A Brief Tutorial on the Numbers That Follow

Text, voice, images, etc, that are stored on computers, or are communicated from one place to another, all require some storage space or communications time.

Computer memory is measured in units called bytes. A byte can usually be thought of as the amount of memory required to store one character of text. Memory needs for voice, images, and other non-text data depend on the type of data and the fidelity required. Memory requirements for sample "files" are shown in Table 1.

Typical Memory Storage Requirements

Page of Text	2 KB
TV-Quality Image	150 KB
Minute of Audio	600 KB
Daily NY Times Text	1 MB
Daily Total FBIS Product	6 MB

Where: B = Bytes, KB = KiloBytes (thousands of Bytes), MB = MegaBytes (millions of Bytes), and later, GB = GigaBytes (billions of Bytes), and TB = TeraBytes (trillions of

Bytes).

Table 1

As can be seen from Table 1, storage needs can vary many orders of magnitude. For that reason, most of the figures that follow will be presented in graphs and charts with logarithmic scales. This has two important side-effects. First, two points or bars may look to the eye to have nearly the same value, in fact may have quite different values—a difference of 1 unit on the 'Y' axis would have a factor of ten difference.

Second, many of the calculations for data costs, transmission costs, etc., are approximations using a number of assumptions. Because of the logarithmic scale, differing assumptions would most likely make very little difference in the appearance of the chart. In other words, if there is a message in the chart, chances are it would come through the same even with different assumptions.

The following sections will describe briefly the current state and, in some cases, projections for the future, of the driving technologies for the system. When that is done, an estimate is made of the costs of each element, and a total cost for the system.

Computing Technologies

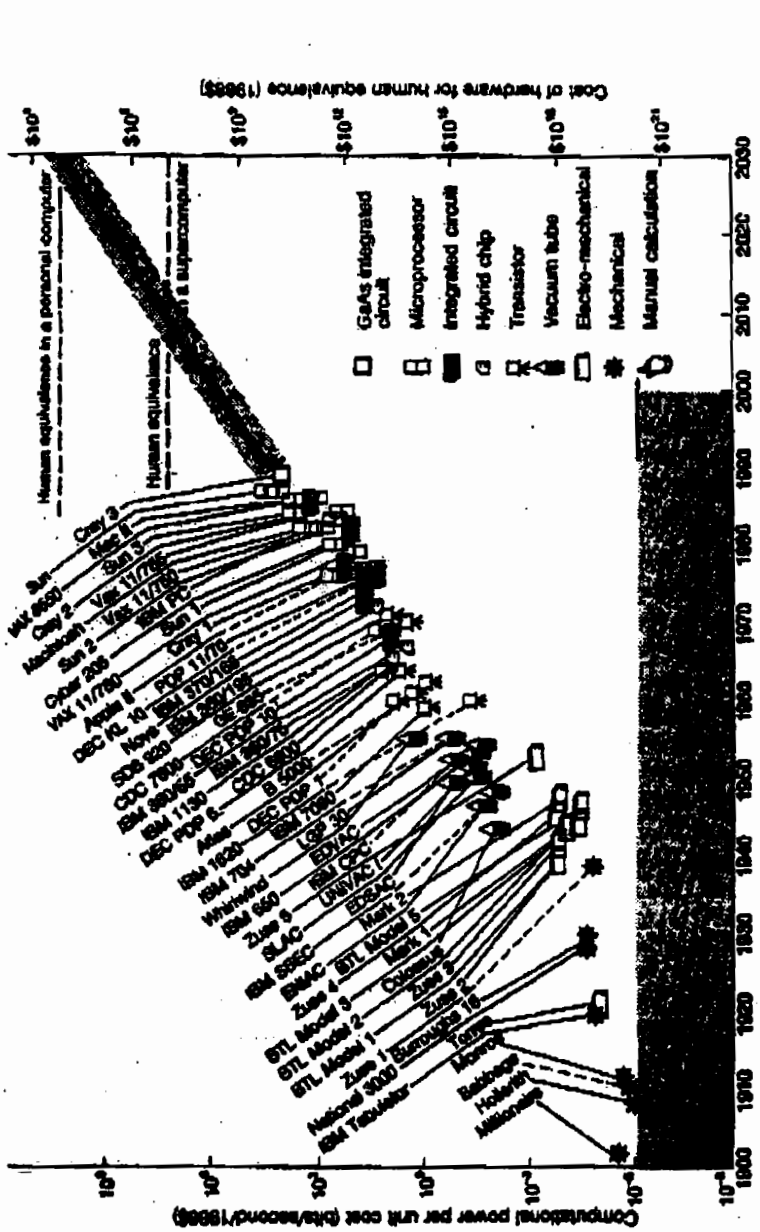
Many studies have charted the progress of computer technology. Figure 3 is fairly typical—it shows that progress, as measured on the logarithmic scale, has been fairly constant since the days of hand calculators. Figure 3 says that computer power doubles every 2.5 years. Put it another way, for the same compute power, the costs drop in half every 2.5 years.

Can this progress continue? DARPA has determined that, not only will the progress continue, but that starting a few years ago, the rate of progress is faster than ever before in history. One of the main reasons is the development of what are called "killer micros"— microprocessors that individually have the power of supercomputers and can be ganged together as one computer.

Information Storage Technology

Information storage costs have dropped commensurately with compute capability. While many technology projections have announced the death of magnetic disk storage systems, they continue to advance and remain cost effective for on-line, high-access-speed systems. Recent trends toward massively parallel disk systems allow system expansion by simply adding more small drives. Built in error checking and correcting protects against loss of data resulting from an individual disk failure.

Figure 4 shows estimated costs for both on-line magnetic storage and archival storage on read-write optical disks.



The evolution of computer power during the twentieth century. Also shown is the equivalent measure of unaided human calculating power and the technological evolution of the computer industry from mechanical devices through electrical machines to contemporary electronic processors.

Figure 3

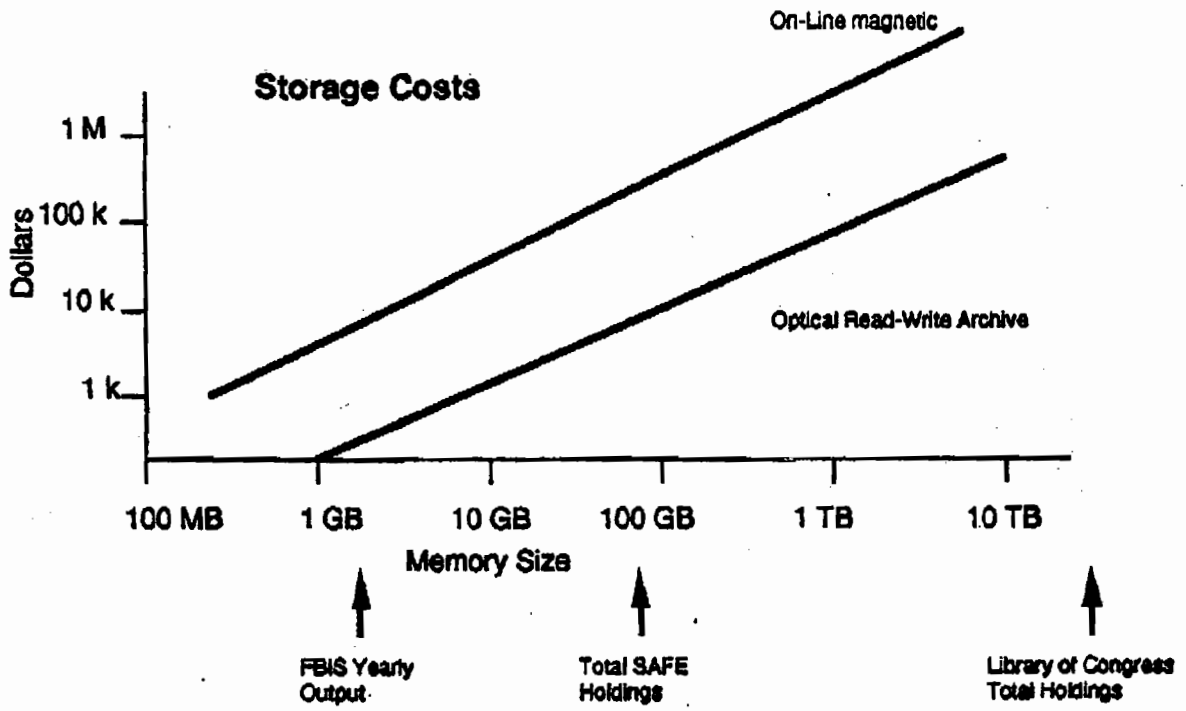


Figure 4

Using the costs from Figure 4 and the storage requirements form Table 1, we calculate the costs of storing different types of files (Table 2).

Costs of Information Storage (cents)

	On-Line Magnetic	Optical Read-Write
Page of Text	0.4	0.03
TV-Quality Image	30	2
Minute of Audio	120	8
Daily NY Times Text	200	12
Daily Total FBIS Product	1200	80

Table 2

Telecommunications Capability

Similar calculations can be made for the costs of transmitting files. Figure 5 shows AT&T's rates for long-haul, wideband communications circuits. These costs have been coming down as the result of widespread use of fiber-optic cables. Projections call for further, drastic reductions in the next ten years.

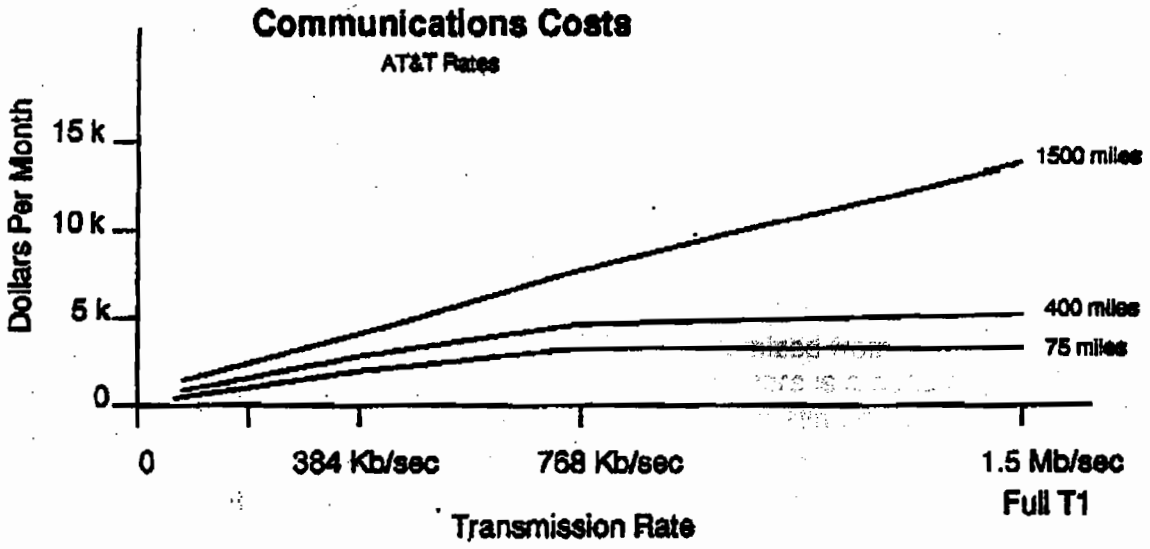


Figure 5

Transmission times and costs for the sample files are shown in Table 3. These assume a 1/4 T1 line (386 Kb/sec) over a distance of 400 miles.

Time and Costs of Transmitting Information

	Time (sec)	Cost (cents)
Page of Text	0.04	0.004
TV-Quality Image	3	0.3
Minute of Audio	12	1.2
Daily NY Times Text	20	2
Daily Total FBIS Product	120	12

Table 3

Software Capability

The system software would need to be assembled and customized from commercial systems. While this would not be an insignificant task, there is a base of examples on the outside that could be used for a start. The UNIX operating system has been designed for this kind of task. Also, the system would be brought up gradually, allowing for tweaking before full demands are placed on the system.

Much user software exists today to get started. Text browser programs are in use today that can manipulate hundreds of megabytes of data—for many foreign languages as well as English. Some good word association profilers and message prioritizers are in use in the CIA today. And CompuServe and Internet have smart "shells" that automatically take a user's request and decide where best to look for the answer.

In sum, software, like hardware, is an area where there will always be a need for new developments. Also like hardware, the commercial world is pushing in the same direction as the Intelligence Community, so we will not have to bear the brunt of the costs.

Data Entry Capability

Data entry has come a long way from the days of the punch cards. Because the computer revolution has affected every aspect of our lives, a great deal of data is already in electrical form, and can be duplicated cheaply. And computer imaging and conversion of printed materials into electrical form, called optical character recognition (OCR), is well along for many languages.

Companies in the information business are selling their data more and more electronically—sometimes, for example encyclopedias, books, and newspapers, very cheaply. Figure 6 gives an estimate of the number of pages of text a dollar will buy today from various types of information sources.

Data Input Cost: What a Dollar Will Buy

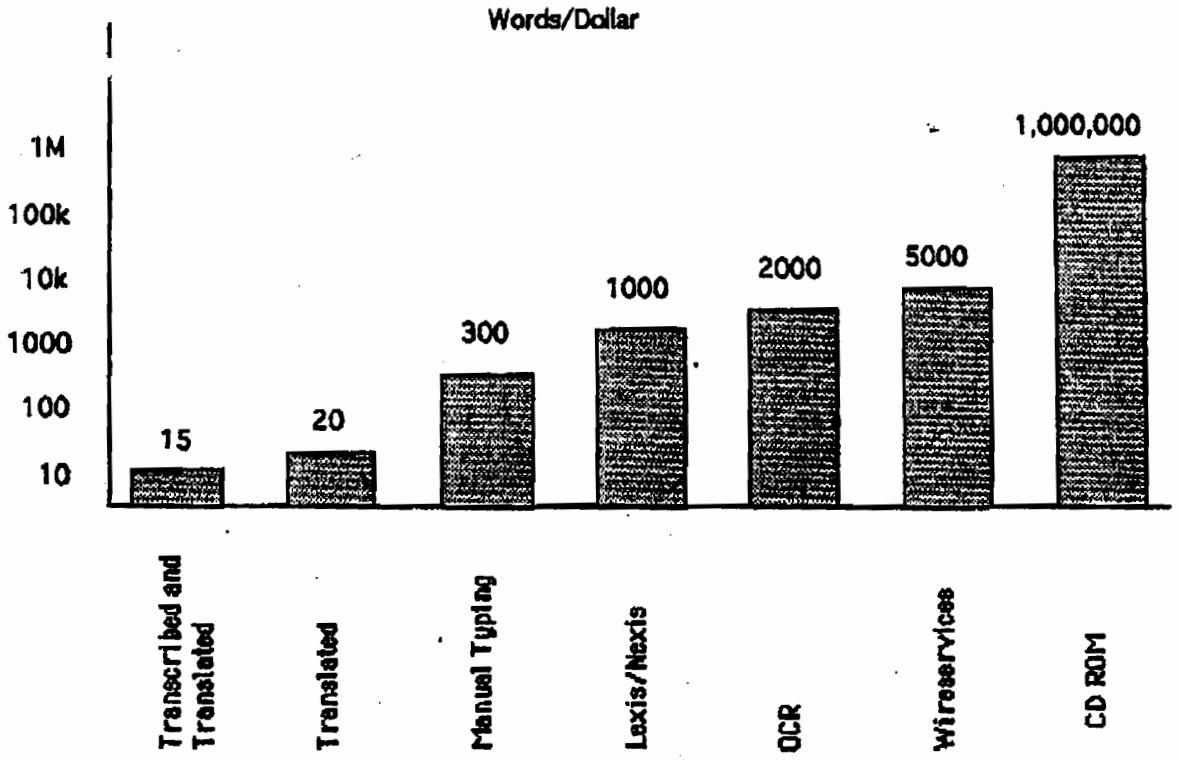


Figure 6

Putting It All Together—Initial Hardware Costs

If the open source network is put in a place where there is already a computer network—computers on analysts' desks and connections to a server—than the open source node or server is the major hardware expense. The cost of the node would depend somewhat on the demand on the node; for example, FBIS's node may need to be more capable that that for State/INR. But, given today's technology, a node would cost an estimated $250,000, about $100,000 of which would be for local magnetic storage. If that node services 250 users, the investment would be about $1000 per user.

This estimate is rough, but in any case, the per user investment would be a small fraction of the $10,000 workstation already on the analysts' desks.

Putting It All Together—Recurring Operating Costs

Assuming that the hardware has a lifetime of four years before becoming obsolete, and that it is replaced at the same cost (but with increased capabilities), the node would be amortized at about $300 per year per user. Making some simple assumptions, per user share of of the telecommunications costs would be about $50 per month. Summing, not counting maintenance, the operating costs would be about $75 per month per user. This compares very favorably with the $1000 per year the Navy charges for connections to its Internet node.

Putting It All Together—Loading the System With Data

Most of the open source data given to the user in electrical form today is from FBIS. The yearly budget for FBIS is $**M, for which they provide about 6 MB of data per day. If we choose to augment FBIS with the following:

Possible Augmentation of Open Source Data (Per Day)

	Bytes	Approx. Cost
15 Newspapers	15 MB	$150
20 Wireservices	20 MB	$600
10,000 pgs OCR of Printed Materials	20 MB	$1000
2,000 pgs Lexis/Nexis	4 MB	$500

Table 4

Then, for $2250 per day ($800K per year), this conceptual augmentation would provide a factor of ten more data to the user.

System Capacity and Expansion Potential

System capacity is difficult to estimate, because assumptions about usage are critical. However, the system appears to be robust. Much of the high capacity postings, such as the newspapers, FBIS reports, and the mass downloads can be done overnight, when users are less involved. If the commonly requested files were broadcast at night in this way, just three hours would be needed to send 100 times the amount of data now sent by FBIS.

Capacity can be increased if needed by adding to both the power of the

computer elements and the communications lines. Communications capacities can be quadrupled simply by going to full T1 lines, rather than the 1/4 T1 lines used as reference here. Moreover, in this study, no use was made of data compression. Because of inherent redundancy in most data, the size of files can be shrunk—at least a factor of 2 for text and a factor of 10 for voice and images—thereby relieving storage and communications requirements, and expanding system capabilities.

In Sum

It appears possible to implement a vision for an open source system for the Intelligence Community that gives analysts far greater access to open source materials and make better use of these materials than today. The envisioned system would be robust, enabling it to keep up with demand and to bring in new services as technology advances.