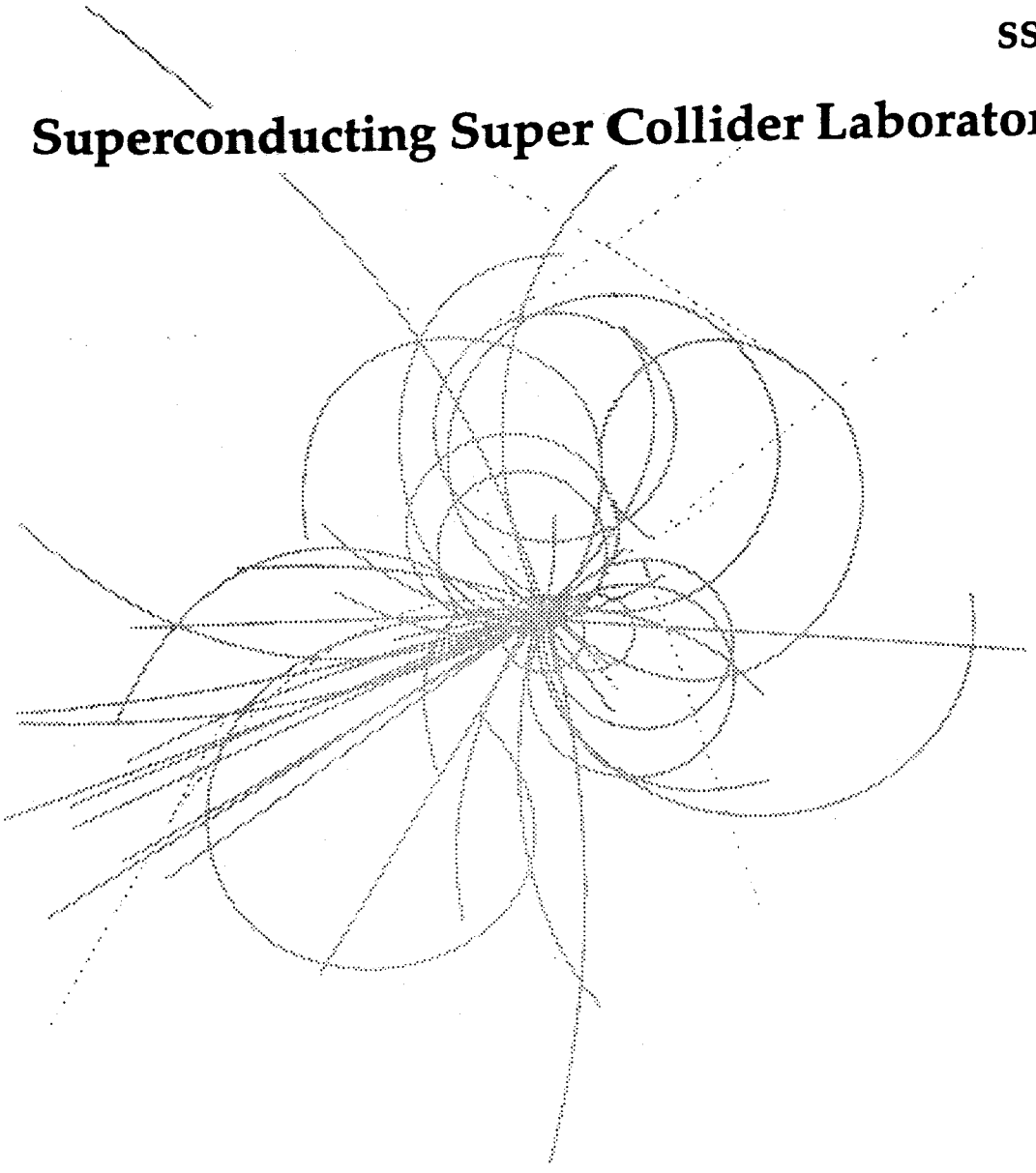
SSCL-390

Superconducting Super Collider Laboratory



Document Format Considerations for a Document Tracking and Storage System

N. Wells, I. Chow, and L. Johnson

March 1991

To be published in Supercollider 3, Plenum Press, New York

SSCL-390

Document Format Considerations for A Document Tracking and Storage System*

Norman E. Wells, Ivan Chow, and Linn. D. Johnson

Physics Research Division
Superconducting Super Collider Laboratory†
2550 Beckleymeade Ave.
Dallas, TX 75237

March 1991

---

*Presented at the 1991 International Industrial Symposium on the Super Collider, Atlanta, Georgia, March 13-15, 1991.

†Operated by the Universities Research Association, Inc., for the U.S. Department of Energy under Contract No. DE-AC02-89ER40486.

DOCUMENT FORMAT CONSIDERATIONS FOR A DOCUMENT TRACKING AND STORAGE SYSTEM*

Norman E. Wells, Ivan Chow, and Linn D. Johnson

Physics Research Division
Superconducting Super Collider Laboratory†
2550 Beckleymeade Ave, MS-2000
Dallas, TX 75237

Abstract: The design and development of the detectors for the Superconducting Super Collider Laboratory (SSCL) will be facilitated by a central system to track and store the design documents and drawings. Collaborators and SSCL personnel need a single system that they can access from a terminal on their desks and use to locate and view the latest version of a document. The SSCL Physics Research Division has developed a prototype system to track documents and drawings and to make them accessible via local and wide-area networks. The prototype is being used to refine the requirements and to develop the procedures for a larger system that will be acquired later. The format for storing documents is a major challenge because there is no interoperable standard. This paper discusses the system requirements and architecture, and presents results of research on standards for storing documents.

INTRODUCTION

The Physics Research Division of the Superconducting Super Collider Laboratory (SSCL) and the detector collaborations require a system to track and store the documents and drawings being used to design and develop the detectors. The system should also be usable for tracking and storing physics papers.

With the geographical separation of detector developers, it is important that a central system be provided so the latest version of a document or drawing is available for a collaborator working on that component or on the interface to another component. The goal is a system that can be accessed via the SSCL local or wide area network so that information is available to physicists and engineers at their desks.

*This work was supported in part by Dr. Tom Kirk and the Argonne National Laboratory for the U.S. Department of Energy under Contract No. W-31-109-ENG-48.

†Operated by the Universities Research Association, Inc., for the U.S. Department of Energy under Contract No. DE-AC02-89ER40486.

Current technology does not support the error-free exchange of information between the many computer-aided design (CAD) and word-processing systems in use at the institutions involved. A limited survey of members of the Solenoidal Detector Collaboration (SDC) found that 20 different CAD and 15 different word-processing systems were in use by the collaborators. The challenge is to develop a system that provides interoperability but does not require the replacement of hardware and software currently in use.

Operating procedures within the SSCL and in some of the collaborations are still being developed. A prototype Document Tracking and Storage System (DTASS) was developed in-house to ensure that a large and expensive system was not acquired until we know exactly how it will be used. The prototype also provides experience that will help specify the complete system to be developed or acquired later. The plan was to get a limited-capability system on line as soon as possible to work out the human/procedural issues and to refine the technical requirements for the larger system to come later.

A format for storing and using documents must be found that will allow system users at the SSCL and associated organizations to access and use the information without buying new hardware and software. Since current technology does not support universal format translations, the Physics Research Division is examining ways to tailor a system to meet its needs.

REQUIREMENTS

System requirements include the following:

- A database management system (DBMS) for tracking the documents and drawings.
- Processing and storage devices for hosting the DBMS and the documents/ drawings.
- Local-area and wide-area networks for communicating with the system.
- Document input and retrieval via menu selection and not by use of a database language.
- Access controls to keep unauthorized personnel out of the system. Controls are also required to allow certain groups to maintain private documents in the system that cannot be accessed by other groups.
- A system to control changes to documents so unauthorized personnel do not modify them without permission.
- The capability to scan and store A- and B-size documents and drawings.
- The capability to add configuration management to the initial system.

ARCHITECTURE

The system architecture is shown in Figure 1. Users log onto the system from a terminal at their desks using the same procedures as those used for accessing the SSCVX1 VAX.

Two Sun SparcStation 1+ workstations, each with 2.5 gigabytes (GB) of storage and tape capability for backups and for mailing large files, are on the network. The Sybase database management system is installed on one of the workstations and contains the index of

all of the documents and drawings in the system. This database (similar to a card catalog in a library) can be searched for subjects, authors, key words, etc. When the desired document or drawing is found, it can be retrieved to the user's workstation.

An input-output workstation comprising a Macintosh IIfx, Fujitsu M3096E/F scanner and QMS PS-810 Turbo laser printer is part of the system. The scanner provides the capability to scan documents that are not received or generated electronically. Hand-written documents can also be entered into the system and stored as a graphic.

Intergraph workstations are connected to the system via the local area network and provide the capability to generate, translate, and store CAD drawings. Users will log onto the DTASS to search for the drawing desired. When a drawing is requested, the database system will retrieve it from the Intergraph server and send it to the requester's workstation via file transfer protocol (FTP).

The document and drawing archive system of the SSCL will be accessible via the local area network and will be able to store documents not in active use.

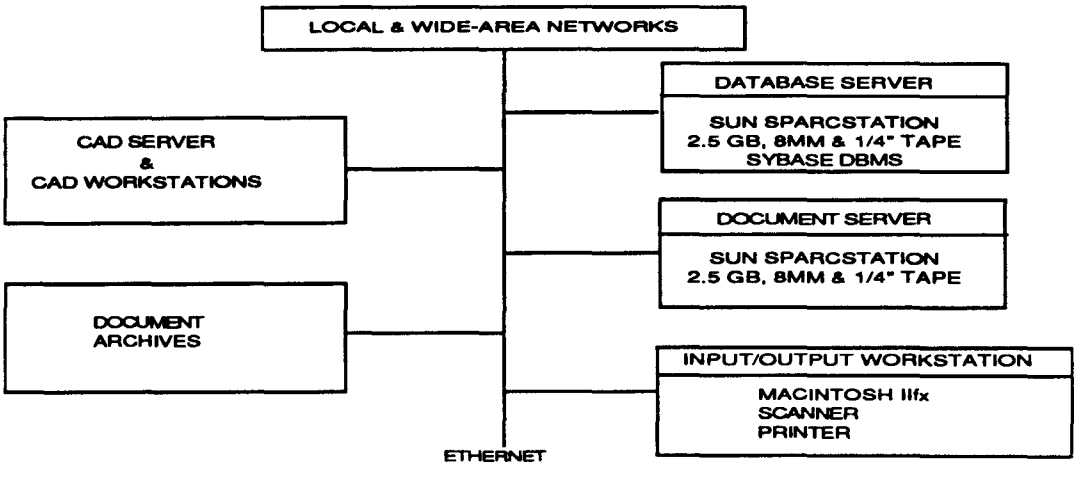


Figure 1. System Architecture.

SYSTEM USE

Users log onto the system via the local or wide-area networks. Choices at the opening menu are:

- *Find/Edit Document*: Users can search for a document by entering the author's name, title, key word, etc. Data similar to that on a library card catalog helps determine if this is the document needed. The current system capability allows the user to transfer a copy of the file to his/her local workstation via a file transfer procedure that is transparent to the user.
- *Enter Document*: Users enter data about the document by filling in the blanks on the screens. The document itself can be transferred into the system by completing additional screen inputs.
- *Utilities*: Users can change their password with this function. Standard reports will be available in the next upgrade of the system.

The system has demonstrated the capability to store and transfer word-processing, spreadsheet, scheduling program, and CAD files. The system hardware, database management system software, and network are all operating. The challenges are to determine the best format for the document files and to develop easy-to-use procedures.

DOCUMENT FORMAT ISSUES

There are many problems to be solved before a user will be able to view a document or drawing at his/her workstation without having to replace existing hardware and software. Current technology does not provide error-free translations between the many word processors and CAD systems in use by the physicists and engineers developing the collider and detectors. The ASCII (American Standard Code for Information Interchange) format for text documents and IGES (Initial Graphics Exchange Specification) format for drawings were initially selected as the standards for interoperability. However, ASCII text does not support storage of the many graphics in physics documents. A search for a better standard is underway.

The interchange of CAD drawings is a problem that may be solved by later versions of IGES. IGES (MIL-D-28000) is being developed as part of the Department of Defense Computer-Aided Acquisition and Logistics System (CALS) that deals with standards for text, graphics, and CAD interchange. At this point, IGES is not yet a standard because many vendors have product-unique extensions that do not convert to other vendors' implementations of IGES. The Physics Research Division is currently concentrating on document format problems and will store CAD drawings in their native Intergraph format for now.

Documents can be accessed by hardware and software different from that used to create them by translating text between different programs, by scanning and storing an image of the page, and by use of a "standard" printer language such as PostScript.

Translators

Translators may be able to convert between application programs if an electronic file of the document is available. This capability (also called a filter) comes with many word-processing programs and can also be purchased separately.

The Microsoft *Word* word-processing program running on the Macintosh was chosen as the initial format for documents because this combination was in wide use at the SSCL. It was believed that getting the system operating with some format was the first order of business; interoperability could come later. Since *Word* is also available for DOS machines and Unix workstations (running DOS for Unix), it was hoped that it could be the interoperable standard for the Physics Research Division. However, research indicated that a physics paper with graphics could not be translated from *Word* running on the Macintosh to *Word for Windows* running under DOS without losing the graphics. In at least this case, a particular word-processing program running on one type of hardware is not interoperable with a word processor with the same name running on another type machine or operating system.

Graphics and Greek letters/math symbols have proved to be major problems in translating between programs. Advertising indicates that translators are available to convert documents between word-processing programs. The authors were able to translate text between word processors and retain most of the formatting, but graphics and equations are a special case for which an easy solution has yet to be found.

Scanning

A hard copy of a document can be scanned and stored in several formats.

- *Raster Image*: A picture of the document can be stored after scanning. This can be done in a graphics format like TIFF (Tagged Image File Format) or PICT (PICTure of the document) used by applications such as *Canvas* or *Optix* for the Macintosh. (*Optix* is a program that controls a scanner and produces raster image files of a document of many pages.) A raster image converted to a graphics application like *Canvas* can be edited or pasted into a word-processing program like Microsoft *Word*. Research indicates that *Canvas* would not hold many pages of images before exceeding the 8 megabytes (MB) of random access memory (RAM) of the Macintosh IIfx. Individual page images can be stored and retrieved in *Canvas* but if the document contains many pages, it needs to be pasted one page at a time into an application that does not try to put the entire document into RAM, such as Microsoft *Word*. A 27-page viewgraph presentation was scanned using *Optix*, transferred to *Canvas*, and pasted together into a *Word* file. This file was 5.3 MB long and required that the RAM allocation to *Word* be increased to handle it. No attempt was made to compress these files since additional user software would be required to decompress them.

  The orientation of images must be considered when they are input. A page input in the portrait orientation with the graphics or text in the landscape orientation must be rotated so it can be viewed. Not all application software programs have a rotate capability.

  A raster image of a document can be stored and retrieved on a central server using a system such as that provided with an *Optix* server. Users with the *Optix* software on their Macintosh can call up and view pages stored in the server at their workstation via the local area network. Since raster image files are large, *Optix* compresses the file for storage and transmittal. The file is decompressed at the user workstation.

The authors have made no attempt to transfer raster images to computers other than the Macintosh.

- *ASCII Text:* Scanned text can be converted to a file of ASCII text by recognition programs such as *Omnipage*. The text output can be put into various word-processing or ASCII formats. However, text recognition is not perfect and many errors should be anticipated. The authors experienced 20-30 errors per page of text. Spell-checkers for text recognition programs are available that claim to clean up most recognition errors. (The authors did not test any of these programs.) Scanning at a higher resolution (300 dots per inch [DPI]) resulted in more errors per page than at 200 DPI. Files of recognized text are much smaller than raster image files: 2-3 kilobytes (KB) versus 50-200 KB. These text files can also be edited or input into other files as if they were created by the user.

  Text recognition programs may not handle graphics. In some programs like *Omnipage*, they need to be "cut" out of the text and handled by a graphics program such as *Canvas*. However, when graphics are pasted into a word-processing program, they may not be edited directly. Greek letters and mathematics symbols may introduce many errors in text recognition and require special handling.

PostScript

PostScript is a device-independent page description language that some printers use to prepare the instructions for printing a page. Files saved in PostScript may be used on various machines and thus may provide interoperability because the documents can be printed or viewed at workstations other than the type on which they were created. However, our initial investigations indicate that there are a number of versions of PostScript, and full interoperability has yet to be demonstrated by the authors. (Additional work is underway by the authors in this area.) Users can convert an existing file to PostScript by invoking the correct command for their computer. Instead of sending the new PostScript file to the printer, the computer stores it on the hard drive, from which it can be transferred to another user.

PostScript files tend to be very large because they define how to create individual letters and graphics elements on the page. For example, a seven-page physics paper (half text and half graphics) was scanned, run through a text-recognition program, and saved in Microsoft *Word*. The graphics were converted to *Canvas* and imbedded in the *Word* document. This *Word* file was 413 KB, but grew to 2.44 MB when converted to PostScript. An 8.5-by-11 in. page containing a CAD drawing of a detector was scanned at 200 DPI into a 449 KB TIFF file and saved as a 144 KB PICT file in *Optix*. The file grew to 287 KB when converted to *Canvas*, but expanded to 1.5 MB when translated to PostScript.

PostScript files can be compressed by about a factor of 10. The software to perform the compression is part of the Unix operating system. Additional work may be required to decompress files on a Macintosh or DOS computer that were compressed on a Unix workstation.

PostScript has the capability to rotate images to correct for storage in the wrong orientation. Panning and zooming capabilities are also provided.

Additional work needs to be done with PostScript to determine its true interoperability. PostScript offers promise for those who need to share files and can handle their large size.

SYSTEM CONSIDERATIONS

These document format issues have a large impact on the total system. Memory is a major challenge—both for storing a document and for accessing it over a network. The resolution of scanners, monitors, and printers may determine whether a document can be read. Manpower to translate or scan documents and input them into the system must be considered as well as the user's time required to actually see the document. Also, the total system should not require expensive licenses at every user's workstation.

Memory Considerations

Since graphics files are so large, a great deal of RAM is required to process them. Even scanning a page of text creates a large file. As previously stated, an 8.5-by-11 in. page scanned at 200 DPI creates a 449 KB file (scanning at 400 DPI results in a 1.7 MB file) in *Optix*. The authors found that the 8 MB of RAM in the Macintosh IIfx used for this research required them to quit applications before another could be opened. This was cumbersome because scanning a document, recognizing the text, processing the graphics, and creating a usable file required the use of *Optix*, *OmniPage*, *Canvas*, and *Word*. The following table indicates the application memory specified by the authors on a Macintosh IIfx for the processing of the documents used for this research. No attempt was made to find the minimum RAM that could be used since it would depend on the documents being processed.

We started with the standard allocation and, when it was insufficient, made arbitrary increases. The following were the final RAM allocations:

APPLICATION	SUGGESTED	USED
SYSTEM	924K	924K
FINDER	350K	350K
TELNET	594K	594K
WORD	512K	3000K
OPTIX	2048K	4096K
CANVAS	700K	3500K
OMNIPAGE	N/A	3000K

Files can be compressed by a factor of about 10. This helps in the storing and transmission of files but takes time for the computer to compress and decompress—at least 30 seconds should be expected. Additionally, the software used for decompression must be compatible with that used to compress the file. This potential threat to interoperability may be resolved because the standards used for fax compression (CCITT [Consultative Committee on International Telegraphy and Telephony] Group III and IV) are also being used for some image compression.

Because the files containing text and graphics are large, access time across the network (and the time is takes to display the file) takes longer than desired. The 27-page viewgraph presentation mentioned earlier was scanned and saved as *Canvas* images stored one after another in a 5.3 MB *Word* file. This file was stored in the Document Tracking and Storage System and retrieved across the SSCL local area network. Transferring a large file like this takes a long time--about two minutes for a computer hooked directly to the Ethernet (ten megabits per second). However, transferring via Appletalk (24.5 kilobits per second) takes about 20 minutes.

Paging through an electronic document to find desired material takes time in the absence of an index or titles on the graphics. Displaying the next page of graphics in the document requires a 10-second wait while it is drawn. A raster image of the document (text or graphics) is just a bit map picture that cannot be searched by a "find" utility in a word-processing program. An index to the document created by the author may be necessary to facilitate finding material if the document is stored as a raster image or PostScript file.

Resolution

The resolution of the various components of the system must be considered before the overall architecture can be determined. Scanning a document at 400 DPI is a waste of memory if the application is going to display it only at 72 DPI. Applications handle displays in two ways: (1) they show each pixel even if only a portion of the page can be viewed on the screen, or (2) they display the entire page on the screen by not showing all the pixels. If the application permits, all of the detail in a document scanned at 400 DPI could be seen by displaying part of the document and letting the user pan around the document to view it all. *Omnipage* displays the page (at least a portion of it, depending on the monitor) at 72 DPI, while *Canvas* and *Optix* show all the resolution available even though the entire graphic may not be seen at once. In *Word*, the user manually sizes a graphic (and loses resolution in the process) to fit on the page, where it is displayed at 72 DPI.

Resolution must also be considered when printing scanned documents. The authors could see no difference in the printed output of a graphic scanned at 200 DPI and at 300 DPI.

It would obviously make no sense to scan an A-Size document at 400 DPI if it were going to be printed at the same size on a 300 DPI laser printer.

The hardware used by the authors had the following resolutions:

SCANNER	200–400 DPI
LASER PRINTER	300 DPI
MONITORS	72 & 82 DPI
FAX MACHINE	200 DPI

Manpower Considerations

The DTASS was designed to allow users to input electronic files directly from their workstation with little effort or system knowledge. In fact, the database fields can be filled in and files transferred to the system in just a few minutes. Everything works correctly as long as the input file can be read by the user who retrieves it. Since there are incompatibilities, an operator may be needed to translate from the native version of the file to the interoperable standard or to scan the document and create a raster image. The system should provide the capability for the operator to input an entire document at one time and not be forced to process one page at a time.

An operator may also be necessary to ensure that the document stored in the system (the interoperable standard file format) is the latest document and matches that of the native/original version. Ideally, users would translate the native version to an interoperable standard, but they may not have the capability to do so. Time to retrieve and actually view the document on the user's workstation must also be considered.

License Considerations

A total system solution must not require all users to buy expensive licenses for their local computers in order to use the system. Public domain software or software included in operating systems such as Unix must be considered. It may also be appropriate to purchase a limited number of floating licenses running on a central computer that allow users to log onto the system, find and view the document they want, and retrieve it with no additional software needed on their machine.

User hardware will certainly need graphics capabilities so that graphics or raster images can be viewed. X Windows will probably also be required. The goal should be to design a system that takes advantage of the hardware that most users have (or are obtaining), and not try to accommodate the terminal with the lowest capability.

SYSTEM OPTIONS

Research to date indicates that the requirements of the Physics Research Division and its associated detector collaborations might be met by (1) a single central server running a word-processor which allows users to view and process documents, (2) a single central server where users can view raster images of documents, or (3) the storage of files in PostScript. Additional options may also be available. The following is only a listing of possible options and should not be considered as the acquisition path that the SSCL or the Physics Research Division is pursuing.

Central Word Processor

A multiuser server with a word processor could be connected to the DTASS and would be operated like the Sybase database management system is now. Users would continue to find documents as they do now, but they would also have a choice of viewing the document. In this case, the central word-processor would manage the viewing and be

controlled by the user from his/her local workstation. This option should result in the lowest memory requirements because documents are stored in text format (with possible imbedded graphics).

If the central word processor were the same one as that used by the SSCL's Technical Information and Publications Department, documents already processed by the Lab would be in the same format. If not, procedures for converting from other formats could be developed to meet the needs of both organizations. Users should be able to convert from the central format to their own if that were necessary.

Central Image Viewer

A central image viewing system could be connected to the database management system to provide the capability for users to view an image of the documents they are interested in. A similar system is in use in the SSCL's Magnet Systems Division using Macintosh computers.

Memory requirements for such a system would be high, and access times over the wide area network would have to be demonstrated to be acceptable. Decompression software would probably be required at each user's workstation.

PostScript

Text and graphics files could be stored in PostScript format in the DTASS and accessed using the same procedures as currently in use. Documents could also be stored in their native format, but they would not necessarily be viewable by all users. Additional work is required to ensure that interoperability can be achieved by all types of workstations. Compatible compression/decompression software also must be demonstrated.

As with the central image viewer, memory requirements and access times would have to be acceptable. An indexing system may also be required to allow users to find desired material quickly.

Other

There may be other options unknown to the authors at this point. The SSCL and Physics Research Division are following the progress of the Defense Department's CALS to ensure that we take advantage of the latest technology in this field.

SUMMARY

The Physics Research Division of the SSCL has developed a prototype Document Tracking and Storage System (DTASS) to provide users with the capability to locate and retrieve documents and drawings. The system is operating and is being used to develop the requirements for a more capable system that will be required later.

A text, equations, and graphics format that can be used on all hardware and with any software has not yet been found. Text has been found to translate between different word-processing systems with little loss of format, but equations and graphics need special attention.

Files can be scanned and stored as raster images of the document. These files are very large. A scanned document can also have text recognized and converted to a text file in the format of several different word-processing programs.

The capability to convert files to the PostScript format allows printing or viewing on devices other than the one that created the file. PostScript files are also very large.

The large size of many of these files creates the most significant challenge to developing a usable system. Large memory requirements make document access and processing slow. The resolution capability of the hardware and software in the system must be matched to ensure that users can actually read the document selected. Manpower requirements to input and retrieve an interoperable document must also be considered in system development.

System options include a central server hooked to the database management system to enable the user to view/process the document on a common word processor or to view a raster image of the document. Conversion of files to PostScript for remote viewing and printing may also meet requirements.

Experience with the prototype DTASS is helping the Physics Research Division of the SSCL develop a usable system. Issues such as developing or locating an interoperable format for documents will be resolved before we acquire a larger system.