AN INTEGRATED INFORMATION RETRIEVAL AND DOCUMENT MANAGEMENT SYSTEM

L. Stephen Coles,
Group Chief Technologist,
J. Fernando Alvarez,
Technical Group Supervisor,
James Chen, William Chen,
Lai-Mei Cheung, Susan Clancy,
and Alexis Wong

Information Systems Integration Group
Institutional Data Systems
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109-8099

ABSTRACT

This paper describes the requirements and prototype development for an intelligent document management and information retrieval system that will be capable of handling millions of pages of text or other data. Technologies for scanning, Optical Character Recognition (OCR), magneto-optical storage, and multiplatform retrieval using a Standard Query Language (SQL) will be discussed. The semantic ambiguity inherent in the English language is somewhat compensated-for through the use of coefficients or weighting factors for partial synonyms. Such coefficients are used both for defining structured query trees for routine queries and for establishing long-term interest profiles that can be used on a regular basis to alert individual users to the presence of relevant documents that may have just arrived from an external source, such as a news wire service. Although this attempt at *evidential reasoning* is limited in comparison with the latest developments in AI Expert Systems technology, it has the advantage of being commercially available.

INTRODUCTION

Today, virtually all large organizations are inundated with data. In attempting to deal with the problem of storage and retrieval from this ever increasing volume of paper, all private and public institutions are exploring new methods for communicating information electronically beyond existing e-mail and Local Area Networks, especially now that the cost of optical storage and scanning technology is becoming more affordable. An additional problem, of course, that we will not consider here, is the graceful transition from the existing infrastructure using a paper data base and manual methods to an all-electronic form of doing business. Clearly, for at least some of the time, both types of systems will have to coexist side-by-side until the all-electronic form gains sufficient acceptance that most people essentially stop relying on paper, and we begin to treat our trees as endangered species rather than as mere commodities. As one example, the Defense Logistics Agency of the U. S. Defense Department, which accounts for about 70 percent of all U.S. Government acquisitions, is planning their CALS Program for Electronic Data Interchange (EDI) so that it will process the equivalent of 200 million sheets of paper per year in an all-electronic form by the end of 1995. The U.S. National Aeronautics and Space Administration (NASA) also retains millions of pages of documents accumulated over the last two decades that it expects to store and retrieve electronically.

This paper discusses the requirements for an intelligent document management and information retrieval system. JPL has now developed a prototype of such a system for NASA Headquarters in Washington, D.C. for the initial storage of 10,000 pages of documents that will be expanded to 1.6 million pages located in a large document archive. It is expected that new document pages will be accumulated initially at the rate of 300,000 pages per year. Access by key words from a full-text search index to both the ASCII form of these documents

as well as their original raster-scanned images must take place in a matter of seconds on any of five standard PC, Macintosh, or UNIX Workstations operating concurrently over a coaxial ethernet LAN that will be upgraded in the next year to a fiber optic network (FDDI). The original scanned documents must always be available, since they may contain signatures or drawings that are essential to the documentation. To implement this system, we have chosen two open architecture 486 IBM-PC servers and a magneto-optical read/writable 88-cartridge jukebox operating over a Novell Netware and a LAN Manager Network. A high-performance scanner feeds pages from an automatic document reader at the rate of 34 pages per minute, compresses them according to a CCITT Group IV standard with a compression ratio of about 20:1, and uses a standard Optical Character Recognition (OCR) software package to convert them to ASCII text. Our experience shows that OCR error rates are quite variable depending on the quality of the source documents and are exquisitely sensitive to such idiosyncracies as the presence or absence of underlining. A high-resolution monochrome scanning station that can capture two full 8-1/2 x 11 pages side-by-side is being used for quality control during scanning and OCR. Color documents are scanned using a separate flat-bed scanner, since the time for scanning a single color document at high resolution typically exceeds five minutes per page. A commercially-available Relational Data Base Management System has been chosen for structured key-word retrieval and the maintenance of user-defined interest profiles.

Figure 1 shows an overview of the hardware configuration of the prototype system. Figure 2 shows the information flow of documents through the system from scanner to optical storage and subsequent retrieval. Figures 3-6 illustrate the additional processing that is needed for indexing files and the arrangement of Directory Structures on the magneto-optical jukebox.

HARDWARE SELECTIONS

The scanner selected was a Fujitsu 3093E (Calera CS100) with a speed of 34 pages per minute, although the Bell and Howell Copiscan II Model 3338 would normally be preferred for high volume work because of its greater capacity (42 pages per minute at 300 dots per inch). The color scanner chosen was the Advanced Vision Research Model AVR-8000/CLX, although any number of other models would have been acceptable including the HP ScanJet IIc or the Epson ES-300c. The compression board selected was from DISCORP, although a competing product from KOFAX would have been acceptable. The Calera WordScan Plus software was selected for Optical Character Recognition. The high-resolution monochrome display selected was the Sigma Design Multimode 120. The scanning workstation platform was a 486/50 MHz client PC with 16 MB of RAM. A Tricord Systems 486 Superserver was selected for the LAN Manager Server, while a 486/50 MHz DX server with 32 MB of RAM was chosen for the Novell image LAN. An HP LaserBank Library Jukebox was chosen as the 88-cartridge (55 GigaBytes) magneto-optical store. A stand-alone magneto-optical drive (680 MB) was attached to the scanning workstation for backup. The SQL full-text document search engine selected was *Topic* made by Verity, Inc. of Mt. View, California. *Topic* provides a comprehensive set of retrieval aids, such as concept-based queries, word proximity queries, synonyms, thesaurus, and so forth. A pair of 16.8 kbaud Courier HST modems from US Robotics were installed in Washington, D.C. and Pasadena running CoSession windows communication software to facilitate remote debugging.

REQUIREMENTS ON SOFTWARE DEVELOPMENT

An important requirement imposed by NASA on DRIMS (the JPL Document Retrieval Integrated Management System) was for hyperlinking raster-scanned images from the original documents to the ASCII text obtained by OCR from those documents after scanning. Thus, if the user identified a relevant document in connection with a structured query search of the full text of the files on the HP-88 jukebox and wished to examine, for example, an original signature specimen or an engineering diagram as it appeared in the original document, he or she should be able to do so quickly (with a single mouse click). Documents were separated at the request of NASA into various categories, including letters (correspondence), memos, proposals, reports, presentations, etc. In addition to scanning in color documents, it was also required that DRIMS have a procedure for handling double-sided documents or even bound documents like books or research papers (for which a wide flat-bed scanner surface is needed). The OCR process must have an accuracy of at least 96 percent

(tolerating at most 4 illegible characters per 100), given an input of high-quality laser-printer text. In order to satisfy the requirement of a user-friendly human/machine interface, DRIMS was implemented in Microsoft Visual Basic under Windows 3.1. Because OCR is such a computer-intensive step in the processing of documents, it was a requirement that after scanning in up to 1000 pages per day of new material, OCR could be carried out during the evenings in a batch mode by passing only the name of a daily file directory with the day's scanned image files. Other requirements, such as user password security and new user set-up by the Data Base System Administrator (DBA) were also implemented in Visual Basic. Using LAN Manager, record-based document retrieval can also be performed efficiently from SQL applications. For example, documents can be rapidly searched by title, date, author, document owner, and so forth without having to search through the full-text *Topic* indices. Document retrieval should be possible from any of the following platforms: PC DOS character mode, PC Windows, Macintosh, and UNIX-based workstations (Sun, DEC, HP, etc.).

RESULTS

The initial users of DRIMS have been the staff of the Office of Space Science and Applications (OSSA) of NASA Headquarters in Washington, D.C. Two stages of data preparation were needed for DRIMS to accommodate the various types of documents. In the first stage, all documents were scanned as TIFF image files. Next, the second stage creates ASCII text files from these image files. Initially, 10,000 pages of documents were scanned and indexed. Documents are scanned-in as either one-page or multi-page documents (an automatic document feeder with a capacity of 50 pages is attached to one of the scanners) independently of their orientation (either so-called "portrait" or "landscape" orientation). A database for DRIMS was created through the combined entries for programs, organizations, events, persons, and other information in reference to the user-defined document types.

CONCLUSION

The transition to a paperless, all-electronic form of institutional work flow will be long and difficult, but the initial stages are now getting underway. Much more experience needs to be obtained with the problem of having humans and not machines correct errors in the OCR process, a labor-intensive, eye-straining, psychologically stressful activity, given that the quality of many of the source documents is quite poor. As one example, faded blue carbon copies of documents typed on a conventional IBM Selectric typewriter 15 years ago, where the registration of characters may not have been perfect, letters may touch one another, and subsequent ballpoint handwriting or official document stamps cut across the text on the document, cause considerable heartburn. Moreover, typing may have been done on a standard *form*, characterized by special boxes outlined by horizontal and vertical lines for input typing, and where the typist was not always careful to ensure that typing never spilled over the boundaries of a box, or distracting overtyping was done to correct typographical errors, and so on. There are also problems with newspaper text, multifonted text (boldface, italics, etc.), and multilanguage/multialphabetic texts. If the document contains tables of budget numbers, for example, the importance of 100 percent accuracy in OCR may be different than if the document contains only conventional text. This process of OCR correction has been known to "burn out" even the most energetic and determined of clerks who are unfamiliar with the domain of discourse. Increasing the resolution of scanning from 300 to 400 dots per inch has been shown to make an incremental improvement in reducing the OCR error rate, but ultimately this approach has diminishing returns. Only the most sophisticated computational linguistics techniques involving not just morphology, automatic spelling-correction, and grammatical correction, but semantic and pragmatic techniques will be needed to reduce the residual OCR error rate down to less than one percent.

Finally, another area for future growth will be provision to incorporate into such a system hooks for multimedia materials, such as voice annotation and full-motion video clips. For this purpose, it will be imperative to comply with the latest standards for color image compression that are currently evolving, such as JPEG, MPEG, DVI/RT (Digital Video Interaction/Real-Time from Intel and IBM), and SGML (Standard Generalized Markup Languages) now being defined by various industry, university, and professional society (ISO) groups.

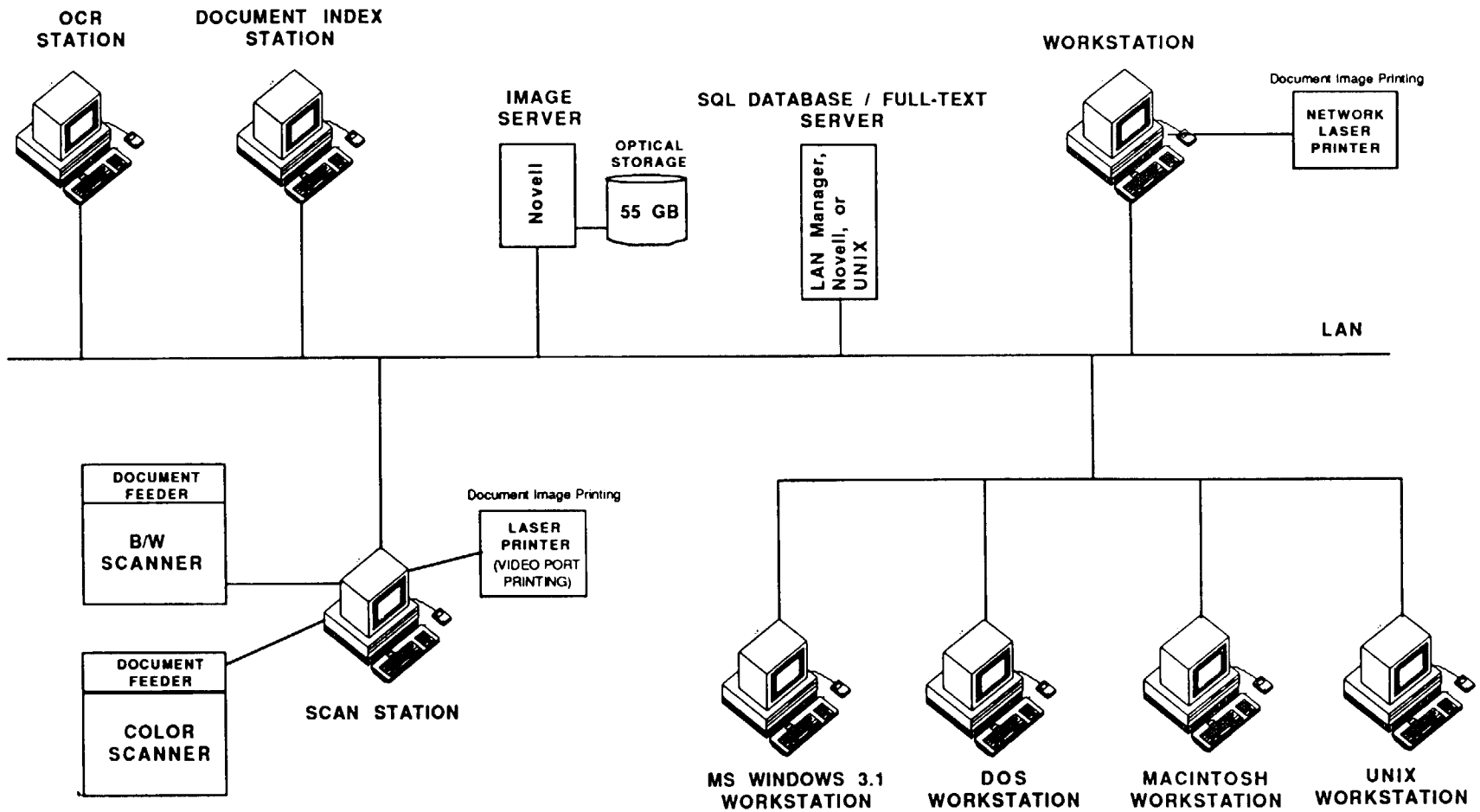
ACKNOWLEDGMENTS

The work described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology under a contract with the National Aeronautics and Space Administration.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

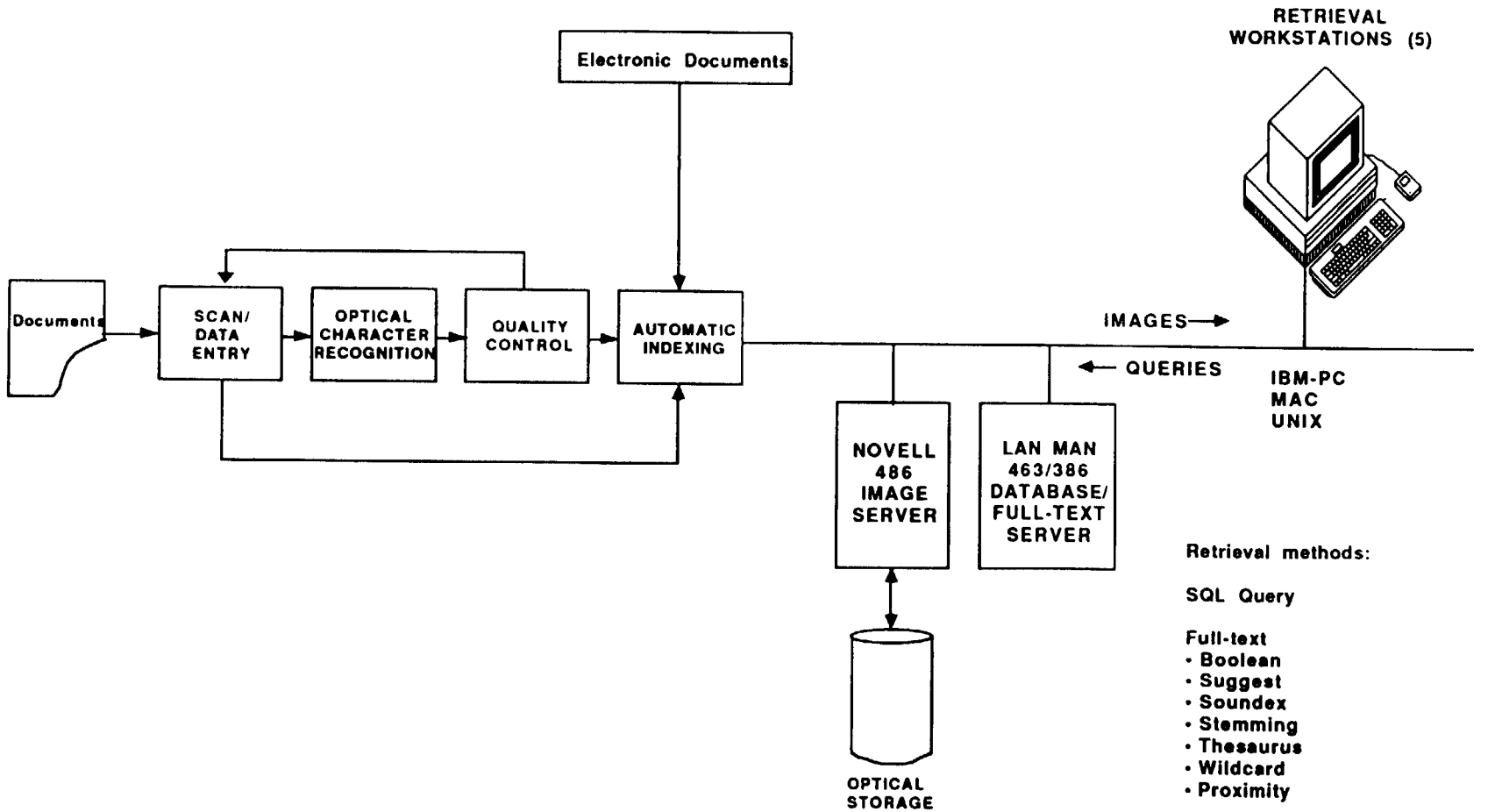
JPL
Jet Propulsion Laboratory
California Institute of Technology

DRIMS HARDWARE LAYOUT



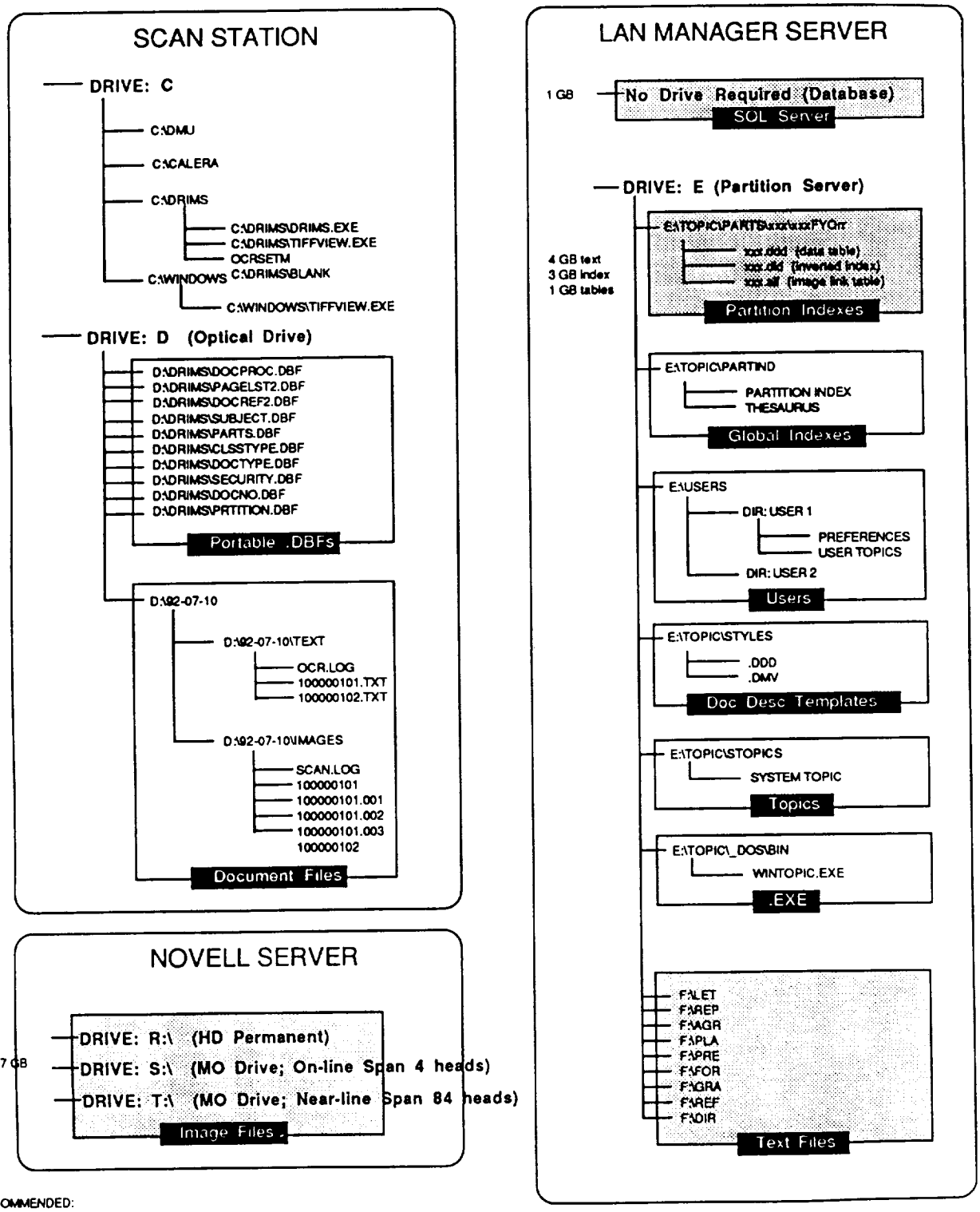
JPL
Jet Propulsion Laboratory
California Institute of Technology

DRIMS DOCUMENT PROCESSING



JPL
Jet Propulsion Laboratory
California Institute of Technology

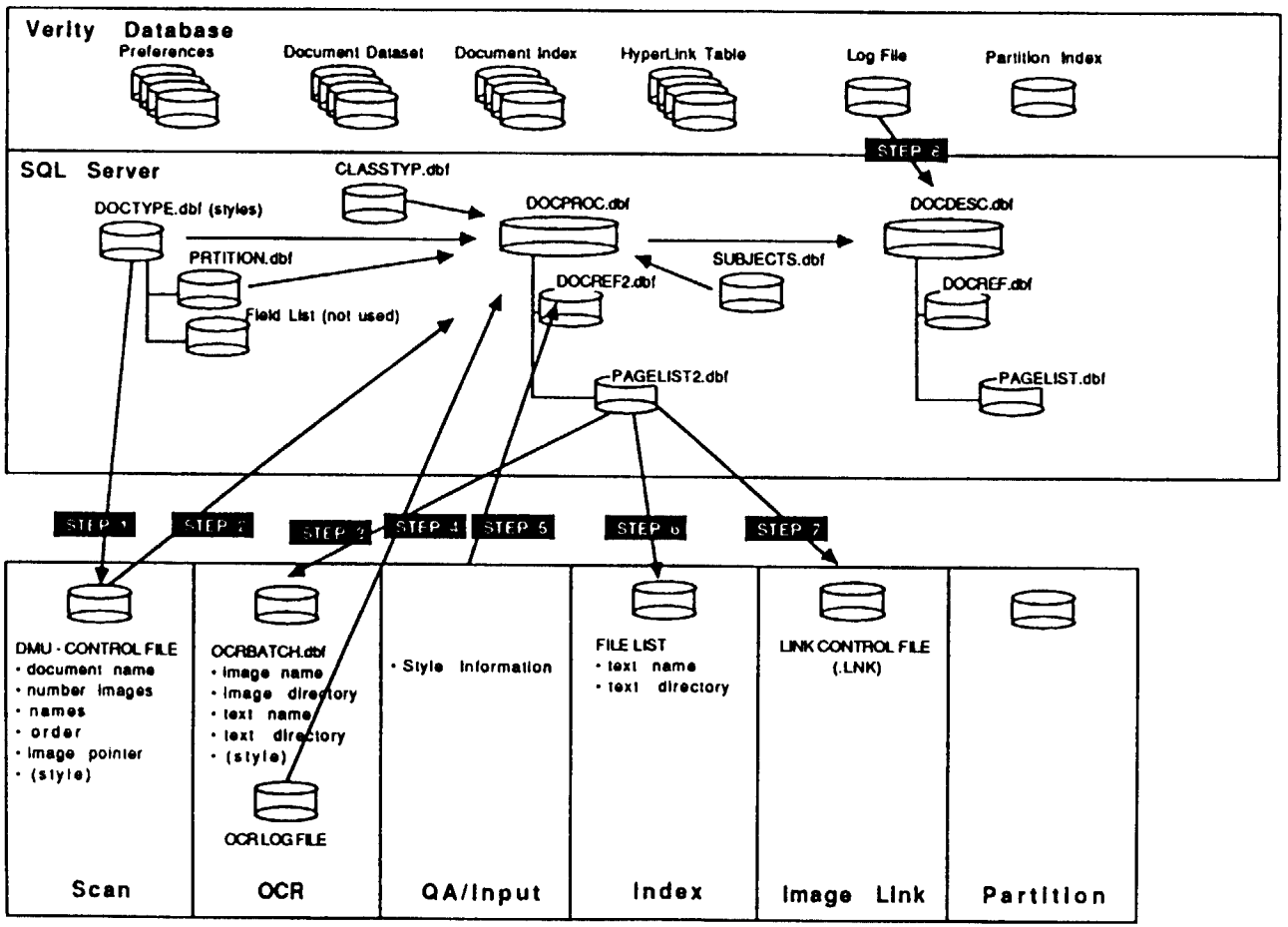
DRIMS Directory Structure



RECOMMENDED:
2-10 MB TEXT PER PARTITION
1-5K DOCUMENTS
2-10K PER DOCUMENT
UP TO 100 LINKS/DOCUMENT
10,000 LINKS/PARTITION

Last Update: 8-21-92
Previous Update 7/7/92
2/18/92

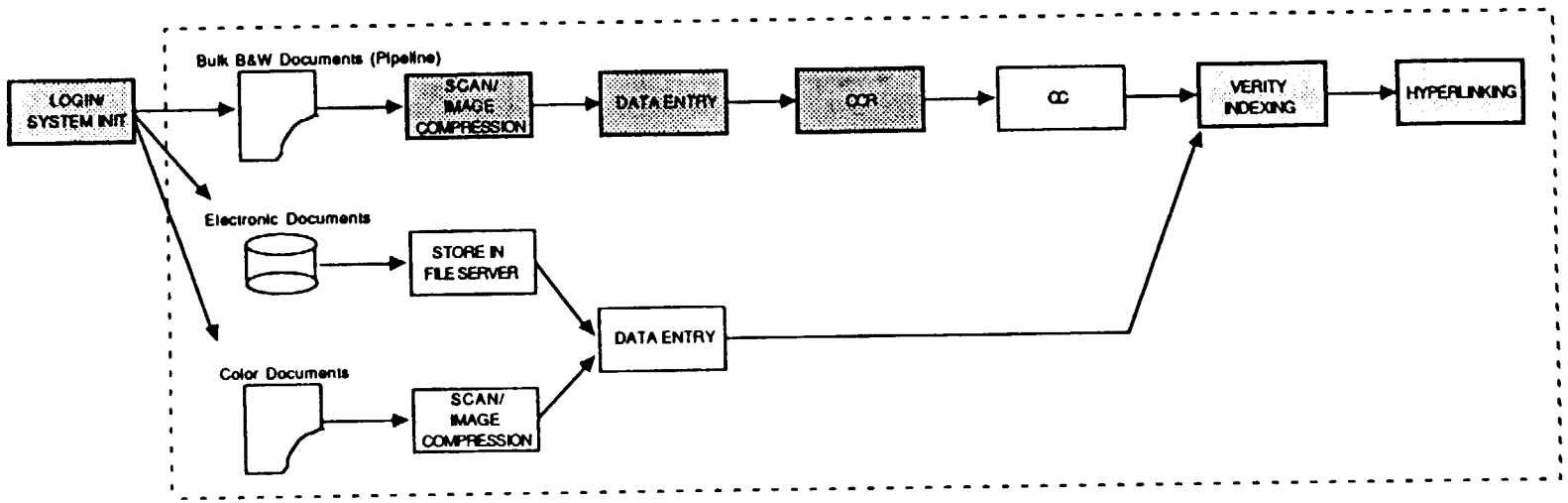
Database Entities



Last Update: 4/22/92
Previous Update: 4/17/92
Created: 2/13/92

DRIMS HIGH-LEVEL DATA PREP FLOW

Level 1

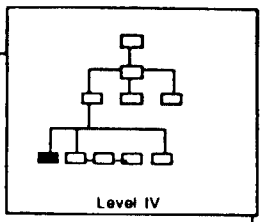


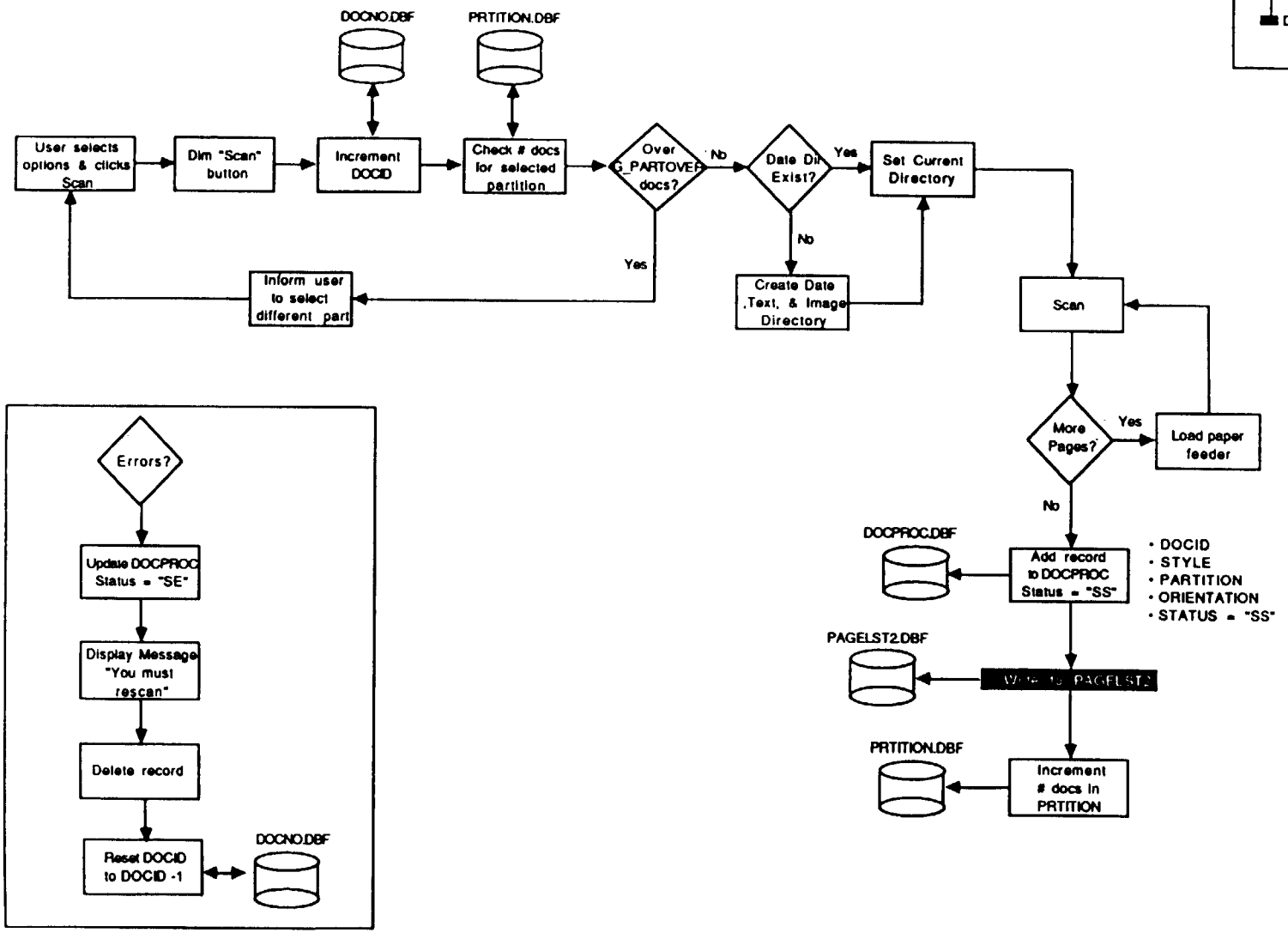
PROGRESS

- UPDATES BEING IMPLEMENTED
- MODULE COMPLETED & DELIVERED
- PROGRAMMING IN PROGRESS
- DESIGN COMPLETED

Last Update: 6/8/92
Previous Update: 6/4/92
Created: 4/07/92

SCAN PROCESS





Last Modified: 7/09/92
Previous Update: 6/11/92
3/24/92