1993

Image processing and optical character recognition : how they work and how to implement them

J. Christopher Reimel

American Institute of Certified Public Accountants. Information Technology Division

INFORMATION TECHNOLOGY DIVISION

Image Processing and Optical Character Recognition

How They Work and How to Implement Them

PRACTICE AID

AICPA
American Institute of Certified Public Accountants

Notice to Readers

This practice aid is part of a series of aids that provide accounting professionals with information about the implementation of a particular technology. These aids are issued by the AICPA Information Technology Division for the benefit of Information Technology Section Members. This aid does not establish standards or preferred practice; it represents the opinion of the author and does not necessarily reflect the policies of the AICPA or the Information Technology Division.

The Information Technology Division expresses its appreciation to the author of this practice aid, J. Christopher Reimel, CPA, who is chief of Information Systems Audit with the New Jersey Department of Labor and a former member of the AICPA Information Technology Research Subcommittee.

Various members of the 1992–1993 AICPA Information Technology Research Subcommittee were involved in the preparation of this practice aid. The members of the committee are listed below:

Everett C. Johnson, Jr, *Chairman*
Kenneth D. Askelson
Paula H. Cholmondeley
William B. Creps
Thomas A. Diasio
Mark S. Eckman
Paul E. Hemmeter
Joseph C. Maida
J. Louis Matherne, Jr.
Robert A. Sellers

Richard D. Walker, *Director*
Information Technology Division

Nancy A. Cohen, *Technical Manager*
Information Technology Membership Section

INFORMATION TECHNOLOGY DIVISION

Image Processing and Optical Character Recognition

How They Work and How to Implement Them

PRACTICE AID

AICPA
American Institute of Certified Public Accountants

1 2 3 4 5 6 7 8 9 0 IT 9 9 8 7 6 5 4 3

Library of Congress Cataloging-in-Publication Data

Reimel, J. Christopher
Image processing and optical character recognition : how they work and how to implement them / Information Technology Division : [J. Christopher Reimel].
p. cm.
ISBN 0–87051–142–4
1. Image processing. 2. Optical character recognition. I. Reimel, J. Christopher. II. American Institute of Certified Public Accountants. Information Technology Division. III. Title.
TA1637.R45 1993
651.5'0285'6—dc20 93–32942
CIP

Table of Contents

Chapter 1. Introduction

Image processing and optical character recognition (OCR) are two technologies that accounting professionals are encountering in their daily operations. These technologies enable companies to become more efficient and to provide better customer service. Although image processing and OCR are not new, their implementation is becoming more widespread as their costs decrease. This practice aid will explain exactly what image processing and OCR can do for a company and how the reader can implement it in his or her company. It will also illustrate the New Jersey Department of Labor's image processing and OCR technology architecture.

Image processing is a method of converting an image on a piece of paper into a binary representation on a computer. The binary representations are then stored, retrieved, manipulated, and integrated with other images and processes to result in an output that can be used at a later date. OCR consists of reading characters on a form (text and numerical) and transmitting this data to a data-processing file or database. Capturing data by OCR is separate and distinct from image processing. OCR data capture can exist without image storage, and images can be stored without OCR data capture. Generally, however, they will be found together.

Image processing technology is similar to the technology used in a fax machine, in which the image on a piece of paper is read into the fax machine and converted into a binary representation. This binary representation is then transmitted over telephone lines to another fax machine, where the binary representation is read and the document is produced. As with fax technology, image processing technology, which was extremely expensive several years ago, is now cost-beneficial for smaller companies. Previously available only on large mainframe computers, image processing technology is now also available for use on personal computers, which are much more powerful than in the past and are therefore able to perform a wider variety of tasks.

The major advantages of image processing and optical character recognition are as follows:

1. Image processing gives administrative employees fast access to a copy of the original document. For many organizations, this fast access can translate into fast customer service, which can give the organization a competitive advantage. In industries such as banking and insurance, where there is little product differentiation, customer service can translate into customer loyalty and increased sales for current and new products and services.
2. Image processing eliminates the cost of administrative time spent searching for lost documents. This is a hidden cost that many organizations underestimate or fail to acknowledge. One of the major advantages of converting manual systems into automated systems is the elimination of time expended on filing, updating, and tracking paperwork. Likewise, one of the major advantages of image processing is the elimination of time expended on filing and tracking paperwork. An image of the original document is available for viewing on a microcomputer screen. This image may be viewed by more than one person at a time. Also, a copy can be produced if a customer or other party requires it.

3. Optical character recognition reduces costly and error-prone manual data entry because it "inputs" the data from a paper form directly into a computer database for processing and report writing. It does not eliminate data entry entirely, however, because some documents are not readable by the equipment.
4. Image processing eliminates the need for storing massive amounts of paper in expensive office locations. If it is believed that the original paper document may be needed in the future, the paper document can be stored off-site in less expensive warehouse space.

Chapter 2. How It Works

Although this book will present a corporate setting to describe a typical configuration in which image processing and optical character recognition are being used, these technologies are also employed by nonprofit organizations and governments. For example, the New Jersey Department of Labor is currently using image processing and optical character recognition to read the employee wage information that employers submit on a quarterly basis with their unemployment insurance taxes. An example of the technical architecture used at the Department of Labor can be found in appendix B.

In many organizations various documents are received from customers or are generated internally. These include purchase orders, receiving reports, invoices, loan applications, and others received on a high-volume basis. Because image processing and OCR require a document to be in a format that facilitates scanning, turnaround documents (for example, utility bills) are candidates for use of this technology, since senders can format the documents to the specifications that the scanning equipment can read when it is received.

In a typical scenario, the customer returns the turnaround document with the payment. The turnaround document typically identifies the customer so that the payment can be credited to the correct account. (Some usual means of identification are customer numbers, account numbers, and mortgage loan numbers.) The document is then scanned by a scanner.

The scanner is connected to a microcomputer through a Small Computer System Interface (SCSI) device. In this example, an imaging system is connected to a microcomputer. One gigabyte of storage has been added through a SCSI device.

As the document is scanned, the scanner creates a Tagged Image File Format (TIFF) and includes in this file a document locator number. This document locator number is extremely important, as it will be used to locate the original document after it has been placed in storage. The index will contain other information that will make the document unique in the system.

For example, a document received from a customer with loan number 1234599 on December 4, 1992, will receive a document locator number of 92339001. The document locator number shows that this document was the first document processed on the 339th day of 1992. A document received from the same customer (loan number 1234599) on December 5, 1992, will receive a document locator number of 92340002. The document locator number shows that this document was the second document processed on the 340th day of 1992. The index on screen might look as follows:

Document Locator Number	Loan Number	Date
92339001	1234599	12-04-92
92339002	8812666	12-04-92
92340001	7733457	12-05-92
92340002	1234599	12-05-92

If you want to retrieve all the documents for loan number 1234599, the index will permit you to retrieve them by using the loan number. The integrity of

the index is extremely important. As shown on the previous page, each document must have a unique identifying field or combination of fields.

In this example, the loan number is unique. Customer numbers and Social Security numbers are also examples of unique fields used in indexes. Because documents are received on various days throughout the year, it will not be possible to retrieve all the documents for a particular loan number, customer number, or Social Security number if the index is not set up properly.

Batches of images are sent from the scanner microcomputer to the archive controller microcomputer and the image server microcomputer. The archive controller microcomputer places the images on an optical disk and sends the index to the central minicomputer. These images are then sent to the OCR boxes.

The OCR boxes read those images that they are able to read and then store the data on a database. The OCR boxes also send those images that they are unable to read to a group of microcomputers, where clerical personnel correct and re-enter the unread characters. These correction and re-entry microcomputers have large-screen monitors to enable the correction and re-entry personnel to view both the original image and the database image at the same time on the screen. This permits the clerical personnel to compare both images and correct the errors.

The OCR box will not let a character go from the image to the database unless it is 100 percent certain that it is readable. The image server will continue to process readable images and reject and reroute nonprocessable images until it has viewed all the documents.

The following are some interesting features of the correction and re-entry software:

- Sections of the image can be magnified.
- Sections of the image can be highlighted with a darkened line used as a ruler.
- The background can be changed from white to black to make the document more readable.

Chapter 3. Conclusion

Image processing and optical character recognition are now cost-beneficial for an increasing number of organizations because of the decrease in price and the increase in power of today's microcomputers. These technologies save money by reducing data entry and by eliminating filing and searching for paper documents. Image processing can be used not only to decrease costs, but also to increase market share by differentiating an organization from its competitors. A service economy, by its very definition, requires that excellent service be provided by the seller. The higher the quality of service, the more competitive an organization's position in the marketplace. Also, many customers are willing to pay higher prices to organizations that provide exceptional service.

Appendix A. Steps in Implementing an Image Processing System

1. Arrange on a flow chart the various systems for which you want to perform image processing. You will be amazed at all the nuances that occur in any processing system. Inexpensive flowcharting software packages are available.
2. Review *completed* copies of the forms on which you want to perform image processing. Determine the various types of errors that can occur when the customer or other third party completes the form. Turnaround documents have the highest read rates. The read rate is the amount of data that can be read, interpreted, and processed correctly without human intervention. For example, a read rate of 90 percent means that nine out of ten data fields were read, interpreted, and processed correctly.
3. Discuss the various types of errors that can occur with the clerical personnel who process them on a daily basis. They will be happy to share their war stories on the variety and frequency of errors. Document these findings to determine which errors you will attack first.
4. Separate the error-free completed forms from the forms that contain errors.
5. Perform your initial tests using the error-free completed forms. Eliminate all the image processing problems and errors that occur in your system while processing the error-free forms.
6. Process the erroneous forms manually until the image processing system is processing the error-free forms to your satisfaction.
7. Perform the cost-benefit analysis for the processing of the error-free forms. For example, the New Jersey Department of Labor has run 80 percent of its forms through image processing and has realized a savings of $5 per form, for a total savings of x dollars. Management is concerned with the immediate dollar savings obtained. Do not ignore this political reality. The system does not have to process 100 percent of the items for it to be cost-beneficial, nor does it have to be 100 percent complete for you to present your dollar savings to management.
8. Process the erroneous forms, eliminating all errors. Refer to the war stories documentation. This is the most difficult, costly, and time-consuming step. Recognize that it may not be cost-beneficial to eliminate all errors committed by the document preparer (customer). Recognize that most systems will always require correction and re-entry personnel.

Appendix B. The Technical Architecture*

The following is an example of the technical architecture in use at the New Jersey Department of Labor. It represents an efficient and effective architecture. However, it does not represent the only architecture available and in use.

Local Area Network

The image/data-entry system is based on the token ring local area network (LAN), which is an open architecture. The token ring allows up to 260 devices to be attached to a single ring, and virtually an unlimited number of devices may be attached with bridges. It is a high-speed LAN that provides excellent performance for image and high-volume file transfer applications, and it operates at a speed of 16 megabits per second with a frame size of about 18,000 bytes. The token ring LAN uses the NETBIOS protocol level.

The token ring network provides a platform allowing the components of the image/data-entry system to communicate. It supports higher-level network protocols such as NETBIOS LU6.2, APPC, SNA, and TCP/IP and operating system software such as DOS, OS/2, UNIX, and AS/400. It is based on the deterministic token passing protocol. Physically it is a star-wired topology with integrated diagnostics. There is no single point of failure, and problem determination is relatively simple. Network management may be local or centralized.

High-Speed Scanner

The high-speed scanner has a maximum throughput of 2,000 pieces of paper per hour. With a vacuum-hold, straight-through paper transport, the scanner is capable of scanning a wide variety of input documents. It is particularly well suited for scanning documents of various thicknesses ranging from 0.002 inch (onionskin) to 0.015 inch (heavy card stock) without adjustments. Document size can vary from a minimum of 2.9 inches wide by 4.1 inches long to a maximum of 9.5 inches wide by 14.5 inches long. The transport mechanism is tolerant of folds, dog-ears, tears, staples, paper clips, and holes.

The scanner's quality control features include a series of enhancement processes to compensate for variations in image density, noise, and background. A monitor displays the images as they are acquired for spot-checking of image quality. In addition, the scanner also includes dynamic threshold detection, auto skew detect, and flat field calibration.

The scanner is configured with an ink-jet endorser capable of printing an alphanumeric text document control number on the documents as they move through the scanner. The endorser operates under program control.

* Adapted from excerpts from unpublished IBM technical material used in response to a request for a proposal to the New Jersey Department of Labor.

Controller

The controller is a microcomputer with 6 megabytes of memory, a 120-megabyte magnetic hard disk, monochrome monitor, a Small Computer Systems Interface (SCSI) adapter, a 16/4-megabit-per-second token ring adapter (which can send data at either 16 million bits per second or at 4 million bits per second), and a 1050-megabyte high-speed SCSI magnetic disk with an access speed of 16.5 milliseconds. The controller software includes bar code recognition capability, which is used for automatic document identification. The scanner attaches to the controller via the SCSI adapter and the controller attaches to the total imaging solution via the token ring adapter.

The controller initiates the scanning process, classifies documents automatically using bar code recognition, and transfers the documents under batch control to the image servers attached to a local area network.

Optical Jukebox

The optical jukebox contains four optical drives and is a high-performance optical disk-handling system for 12-inch optical disks. The system can have up to five drives and as many as 144 optical disk cartridges, for a total of 288 gigabytes of information.

The jukebox is attached to a microcomputer controller with 4 megabytes of memory, a 121-megabyte hard disk, a monochrome display, a 16/4-megabit-per-second token ring adapter, a 1050-megabyte high-speed SCSI magnetic disk with an access speed of 16.5 milliseconds, and a SCSI adapter.

Data Entry and Index Server

The minicomputer serves multiple functions in the overall system design. It provides the interface and storage media for indexing of images. As the index server it contains the locator database. The index is provided to the microcomputer as an advanced-peer-to-peer communication transaction on the LAN. The index entries in the database will include information such as unique identification number, document identification number, physical location on the optical jukebox, and status. It is this index that provides the path for rapid and reliable retrieval of images as they are needed for daily operations. The status indicator determines how far a document has progressed through the system.

Image Server

The image server is attached to a LAN and provides work-flow management, image storage, and image retrieval services. The server consists of a microcomputer with 8 megabytes of memory, a 115-megabyte magnetic hard disk, a SCSI adapter, a 16/4-megabit-per-second token ring adapter, a color monitor, a 1050-megabyte high-speed SCSI magnetic disk with an access speed of 16.5 milliseconds, and an OCR subsystem with a SCSI interface.

The server uses the advanced-program-to-program-communication protocol to communicate to the data-entry and indexing server and to other image servers. This enables multiple servers to be connected over the LAN. This feature also enables remote servers to communicate over the same data network used for host-based communications.

A utility allows users to develop a data-entry screen for a structured form processed through the imaging system. This utility allows the user to specify the type of data that will be entered into each field, to specify the order of data-entry fields, and to control the appearance of the fields. It has table functions and type

checking of data to enhance the quality control of transactions. The data-entry window can be modified at any time; changes are distributed over a network from a departmental server to workstations. The utility fields are used to specify OCR retrieval from images.

The OCR subsystem attaches to the image server with a SCSI interface and processes at speeds up to 250 characters per second. Depending on the amount of information on a form, the system can process as many as 15,000 forms in a twenty-four-hour period without operator intervention. The text recognition system intelligently extracts characters from documents by recognizing typewritten, typeset, and computer-generated print.

A utility allows users to define rules for the electronic movement of documents to any location in the network. Authorization and protection codes can be defined to guide the document routing.

Workstation

An imaging workstation provides a side-by-side display of image, text, and mainframe terminal sessions in a windowed environment. Windows are displayed on a high-resolution 1600 x 1200 pixel 19-inch monitor and may be moved and resized as desired.

The workstation consists of a microcomputer with 5 megabytes of memory, a 60-megabyte magnetic hard disk, a 19-inch monochrome monitor, an image adapter with memory kit, and a 16/4-megabit-per-second token ring adapter.

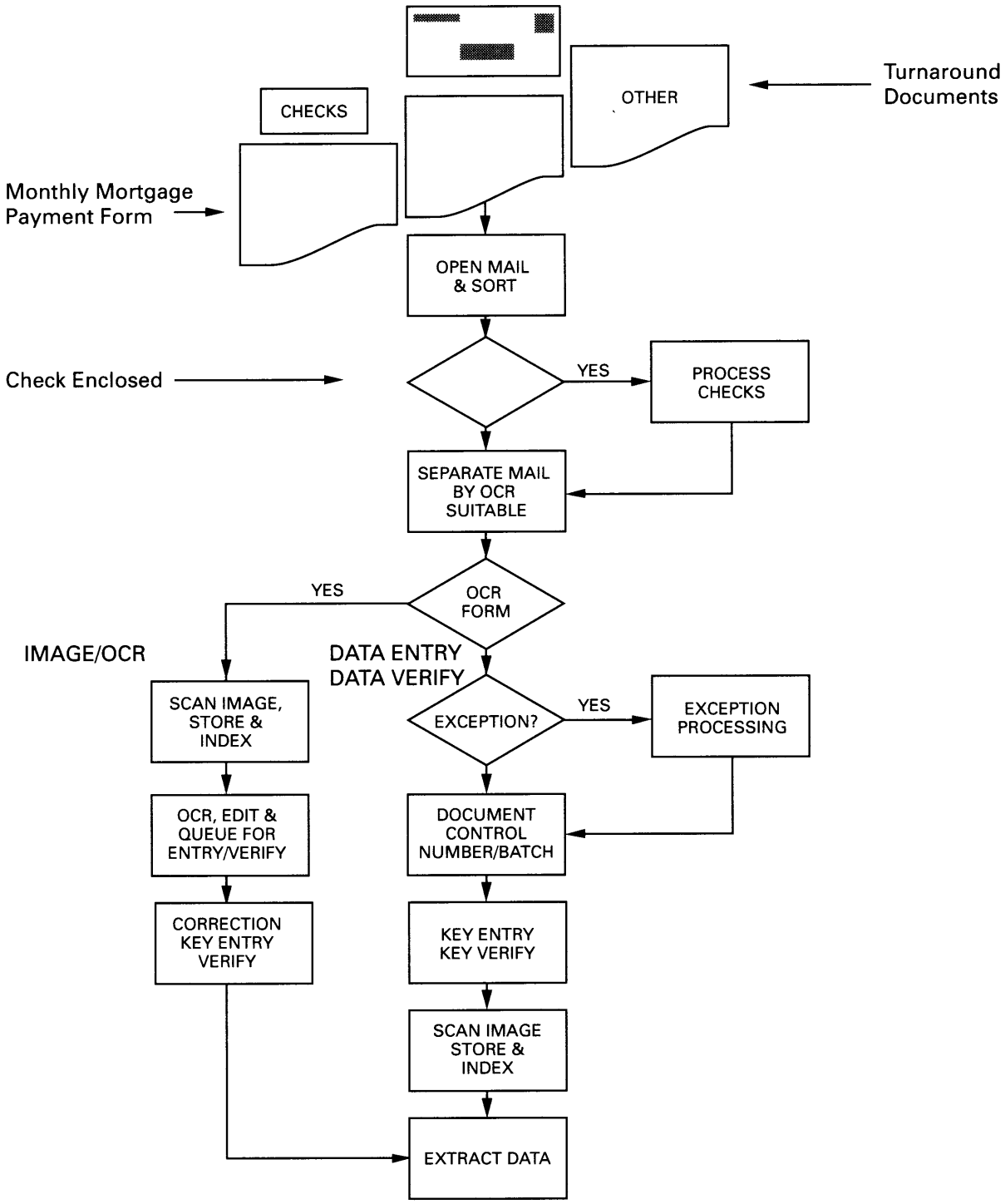
The workstation provides all the material typically used in an electronic data-entry process. The paper is displayed as an image on the screen next to the data-entry fields. As the operator tabs to each field, the corresponding area of the image is magnified and displayed directly below the data-entry field for easy and clear key entry. When all the data has been keyed from the current image, the next image is displayed and the data-entry fields are prepared to receive the new data.

Other features in addition to data entry are provided to enable a user to retrieve images and connect to host applications in order to respond to a customer inquiry. All of the information can be displayed on a single screen, thereby saving time and improving customer satisfaction. The workstation also includes typical desk accessories such as a calculator and a notepad.

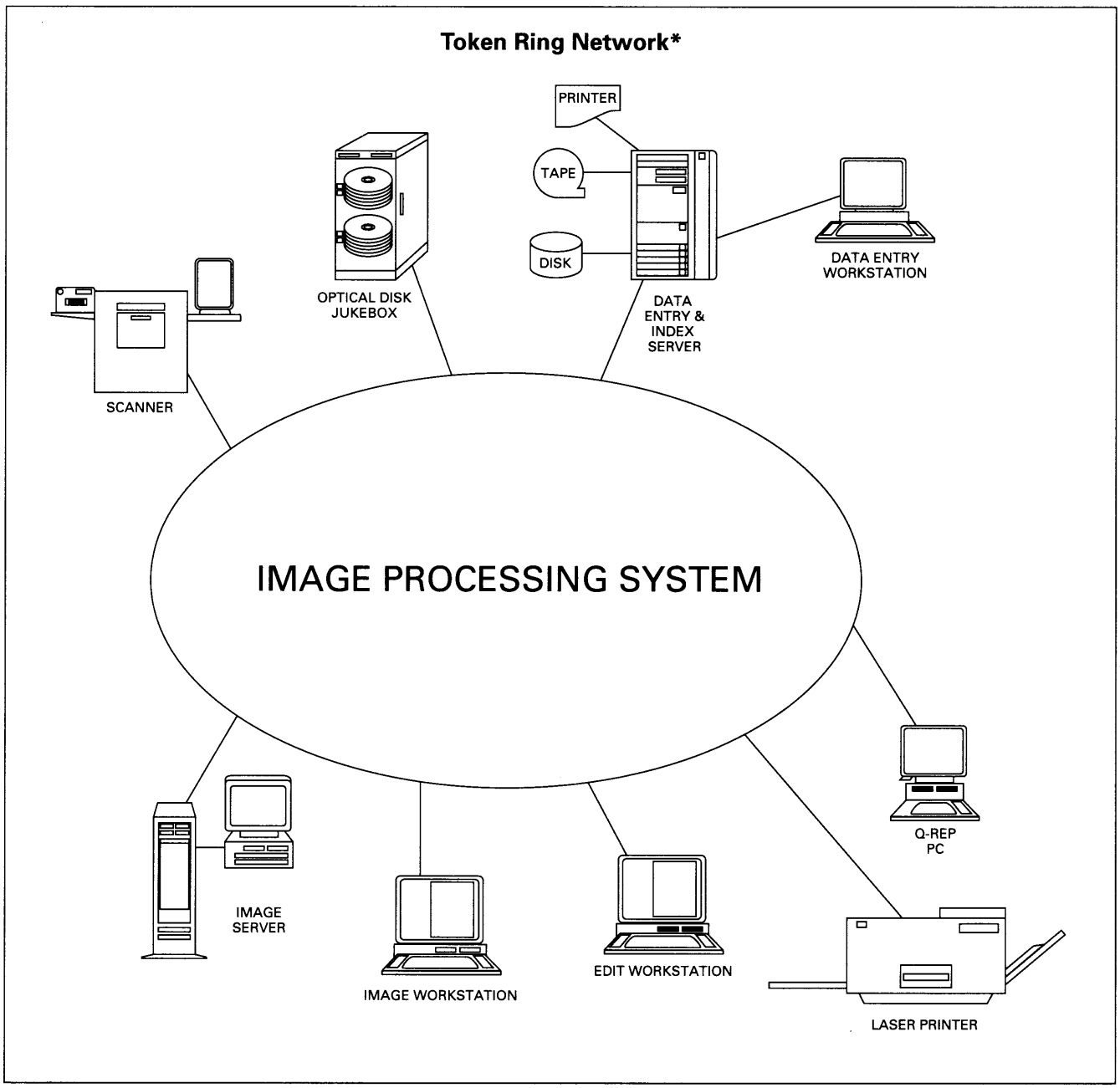
The Minicomputer

The minicomputer collects all data whether it is provided by the scanning system or by direct data entry. The data-entry system is totally independent of the imaging system. The data-entry system residing on the minicomputer provides many ways in which to edit data and ensure its integrity. Some of the data validations are field checks, digit calculation checks, and reasonableness checks. The system also provides for batch balancing and arithmetic calculations, and for program calls to further validate the unique identification field. The minicomputer is configured to accommodate up to 120 devices (terminals and printers) with no need to add controllers.

Overview of System Flow*



* Excerpts from unpublished IBM technical material used in response to a request for proposal to the New Jersey Department of Labor.



* Excerpts from unpublished IBM technical material used in response to a request for proposal to the New Jersey Department of Labor.

Appendix C. Image Compression Techniques

Storing a page of text can require up to 4,000 bytes. When this same page is scanned, it will require approximately 35 million bytes. Most pages have a lot of space without print. This space is known as white space or redundancy. The process of eliminating this unprinted space is called compression. The compression ratio is the ratio between the number of bits in the image before the compression occurs and the number of bits after the compression occurs.