A STUDY IN APPLYING
OPTICAL CHARACTER RECOGNITION TECHNOLOGY
FOR THE FOREIGN BROADCAST INFORMATION SERVICE
FIELD BUREAUS

by

William V. Stine

Project report submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science

in

Systems Engineering

APPROVED:

Benjamin S. Blanchard, Chairman

T-C Poon

Joseph Lewis

January 1993
Blacksburg, Virginia

LD
5655
V851
1993
S756
c.2

A Study of Applying Optical Character Recognition Technology for the Foreign Broadcast Information Service Field Bureaus

by

William V. Stine

B.S. Blanchard, Chairman

ABSTRACT

The Foreign Broadcast Information Service (FBIS) collects and disseminates world-wide open-source information for the U.S. government through a collection of 17 field sites, or bureaus, located in cities around the world. Several bureaus collect a large amount of English-language material that is manually rekeyed into a computer database. Since this is a labor-intensive process that prevents some bureaus from meeting their processing requirements, FBIS is interested in applying a system that would make this process more efficient.

From analyzing the requirements for an improved text-entry system and evaluating several alternative solutions in terms of cost and feasibility, a design approach using Commercial Off-The-Shelf (COTS) Optical Character Recognition (OCR) technology is recommended. The technical requirements for using OCR systems at FBIS field sites is presented along with evaluation techniques for choosing cost-effective COTS OCR products. Finally, the requirements for testing alternative OCR system designs under field operating conditions is included to determine the specific range of printed materials effectively processed by OCR.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

FBIS collects and processes open-source foreign media from field sites, called bureaus, located in cities around the world. At several bureaus, a large amount of English-language print material such as newspapers is processed by the manual-intensive method of rekeying articles, word for word, into a word processor. With significant increases expected in the amount of this material required to be processed, FBIS is interested in investigating ways to make the processing procedure more efficient.

Figure 1 outlines this report's objectives. This report describes the problems with the current manual text-entry method and the specific user requirements for a more efficient text-entry system. Several alternative solutions are evaluated in terms of cost, effectiveness, and technical feasibility. From this analysis, an Optical Character Recognition (OCR) system design is recommended as the most feasible approach. The operational and technical requirements for an OCR system are described, where system components consist of Commercial Off-The-Shelf (COTS) products. From these requirements, a cost/benefit analysis method is presented for selecting the most cost-effective COTS OCR system designs. Finally, a recommended field test and evaluation plan is included for testing proposed systems under field conditions. Based on FBIS requirements, a proposed schedule is presented to implement OCR systems to designated bureaus by 1 January 1994.

This report consolidates user and technical requirements data from FBIS headquarters and field bureaus. From studying FBIS needs and selecting an OCR design approach, this report presents a method for determining those bureaus where OCR systems are required and the specific sources of printed material where OCR processing is effective.

Before analyzing the system deficiencies and needs, a brief background and overview of the FBIS mission, bureau technical systems, and bureau operations is presented

1.1 Mission of FBIS

The mission of FBIS is to collect, select, process, analyze, and disseminate open-source foreign media to various US government agencies. Seventeen bureaus located around the world collect and translate international publications, documents, radio, television, and newspaper press. The English translation of this

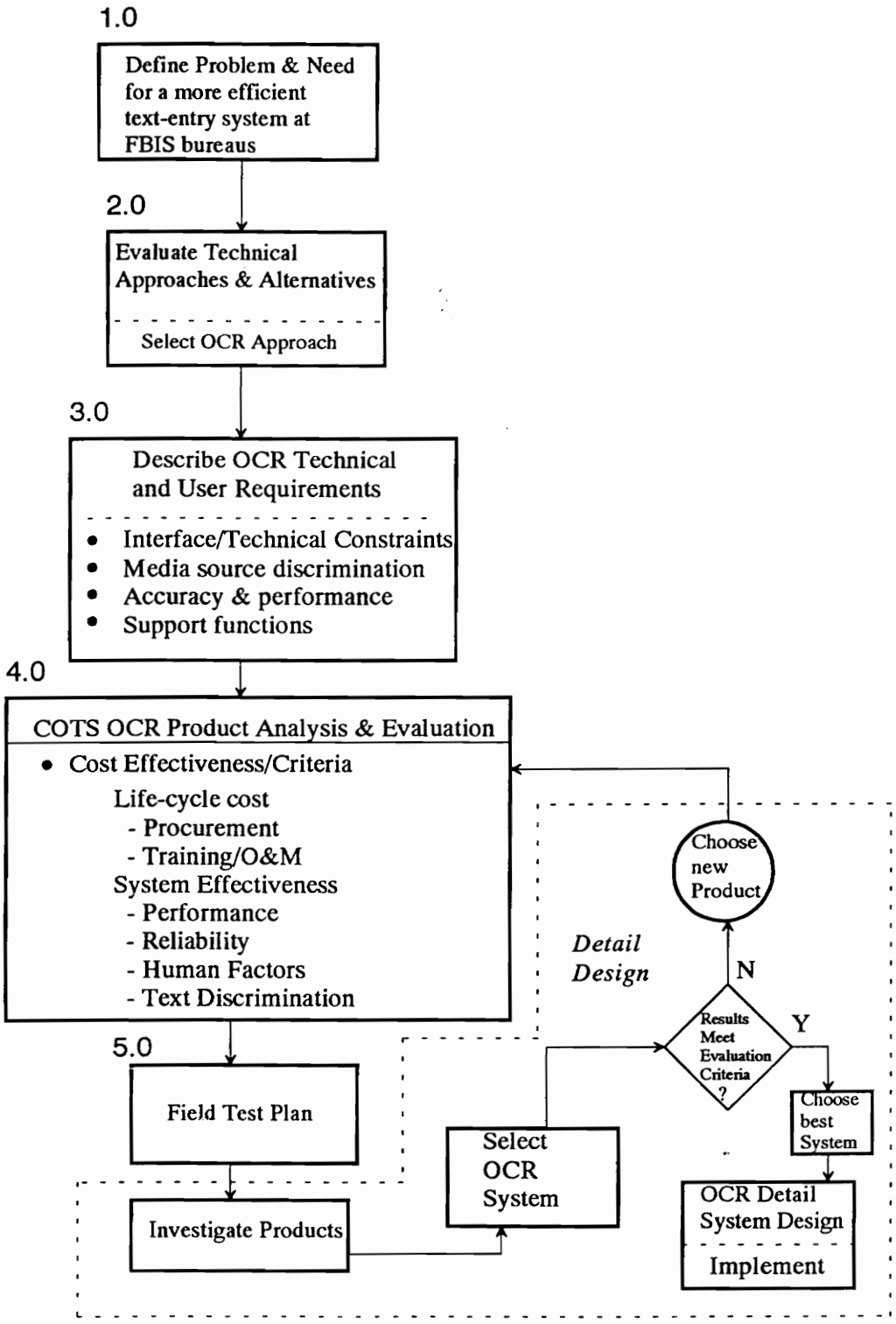

Figure 1. Scope and Flow of Investigation

information is transmitted to the United Stated for distribution to U.S. policy-makers and government agencies.

It is the objective of FBIS to provide U.S. officials with timely and accurate information on world-wide political, economic, military, and scientific items of national interest. Every weekday, eight volumes of material called Daily Reports are published. Each volume represents a different area of the world and contains the verbatim translation into English of the original foreign media items.

The FBIS collection and dissemination system allow items, or "messages," to be prioritized and transmitted to a wide range of customers, called "consumers". For example, high-priority items that occur in a crisis situation can be transmitted directly to interested consumers. With the increasing amount of unrestricted foreign media and global competition, the importance and volume of open-source information is expected to increase for the foreseeable future.

1.2 Field Bureau Operations

Each bureau collects specific sources of information based on established geographic coverage areas. As a result, the size, objectives, and operational procedures of each bureau vary according to their specific collection requirements.

Each bureau collects most of their information from Foreign Press Agencies and radio/TV broadcasts. Most of this information is received automatically via antenna, or from subscriptions. Radio broadcasts are pre-selected so bureau personnel called "Monitors" know exactly when broadcasts (such as news reports) occur. The Monitor translates and types in selected broadcasts, exactly as they are received, into the computer system as text items.

The types of hard copy English-language print material collected by bureaus include newspapers, facsimile, periodicals, and magazines. Selected items are retyped into a word processor. The selection, coverage, and scope of this information is very broad and depends on factors such as local geographic areas, available resources, receivability, availability, and priority.

The main technical systems of a bureau are divided into collection and processing components as shown in Figure 2. Foreign media sources are collected at each bureau from antennas, satellite dishes, radio receivers, and telephone lines. The information is entered, processed, and transmitted to consumers through the use of a Personal Computer (PC)-based Local Area Network (LAN) and communication system. Items are transmitted to local consumers within the particular region, or back to the United States to be published.

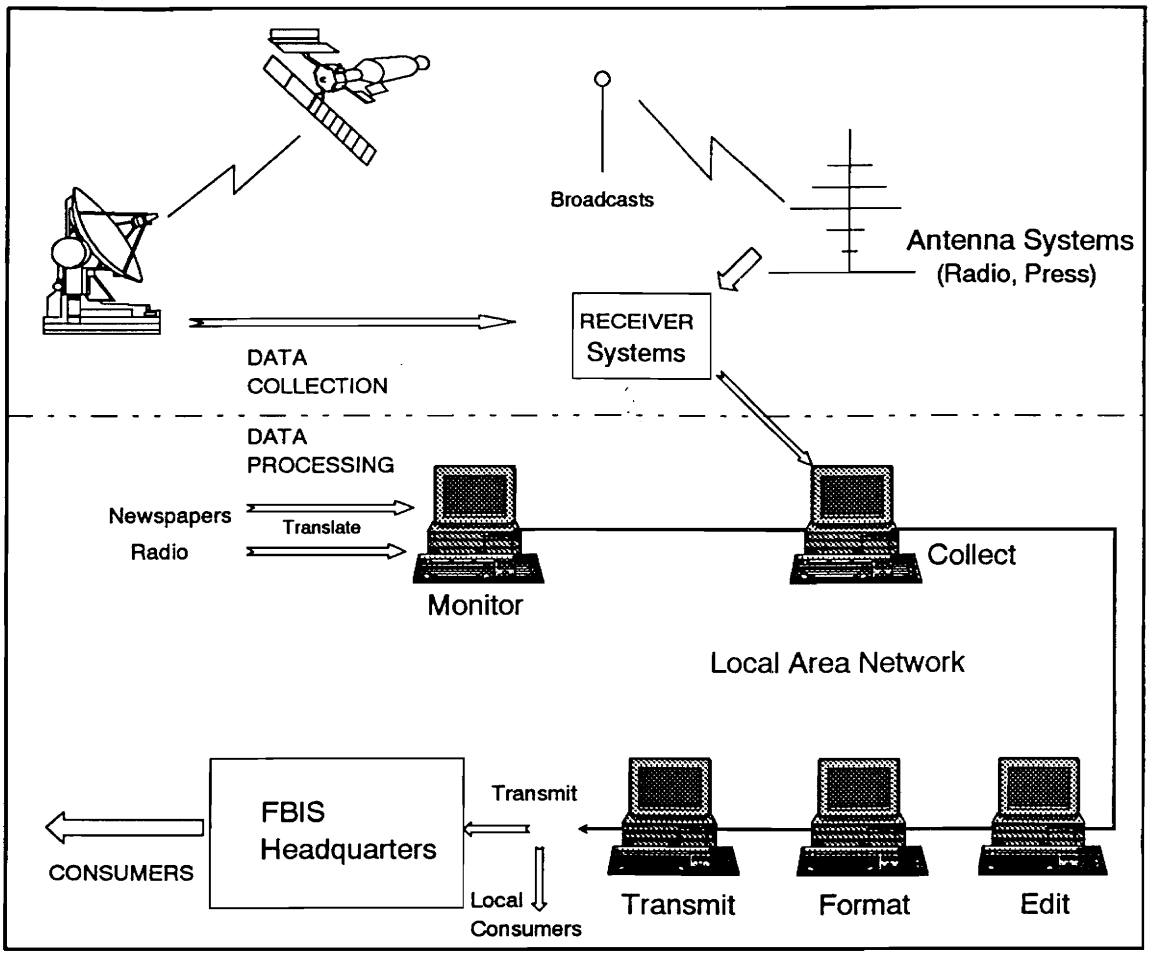

Figure 2. FBIS Bureau Technical Systems

All bureau collection, processing, and communications support is provided by the Engineering Support Group (ESG) located at the central FBIS collection and dissemination site located in the United States (FBIS Headquarters). ESG evaluates technical requirements, performs engineering studies, designs and implements installations and upgrades, and provided daily operational and maintenance (O&M) support.

1.2.1 Bureau Collection System

As shown in Figure 2, FBIS bureaus receive most of their information by electronic means. For FM and AM broadcasts, radio receivers are tuned to specified frequencies, recorded, and played back when needed. Many of the Foreign Press Agency news broadcasts are connected directly to the processing

system through the use of 19.6 Kilobyte serial lines connected via modems to antenna receivers.

1.2.2 Bureau Information Processing System

The computer processing system shown in Figure 2 supports all processing tasks from the input of an item until it is transmitted to consumers. Although each computer is attached to the same network, each performs specific operational functions through the use of specially developed software. Figure 3 shows the computer processing components of a bureau, including the various workstations and their organization within the bureau. The workstations shown perform the following specific processing tasks:

- MONITOR - Data entry and translation of media source items
  Totals about 75% of the bureaus' computers
- EDITOR - Selection, editing, and delegation of source items
- FORMAT - Preparation and formatting of messages for transmission
- TRANSMIT - Transmission of messages to appropriate consumers
- PAC - Press Agency Collector, collection of press agency items
- File Server - Storage of all shared data, resources, and applications
- TECH/ADMIN - Bureau Technician and Administrative support
- IDD - Backup transmission of items via telephone lines

1.2.2.1 Computer Hardware and Software Platforms

The current computer system hardware baseline at FBIS bureaus consists of XT and AT-class IBM-compatible computers connected by a LAN file server using the Novell Netware Operating System. The bureaus are being upgraded to a TCP/IP Ethernet LAN and standard 486-processor computers to take advantage of current industry standards, increase processing speed, and to provide a stable platform for expandability.

The processing software used on each computer is referred to as the Field Automation Segment (FAS). Although most of the FAS software has been specially developed for bureau use, it includes various COTS software for word processing and task-switching functions. For example, FAS uses a word processor called XyWrite for all text processing. It provides not only text input and editing, but unique output formatting capabilities that are required in the FBIS

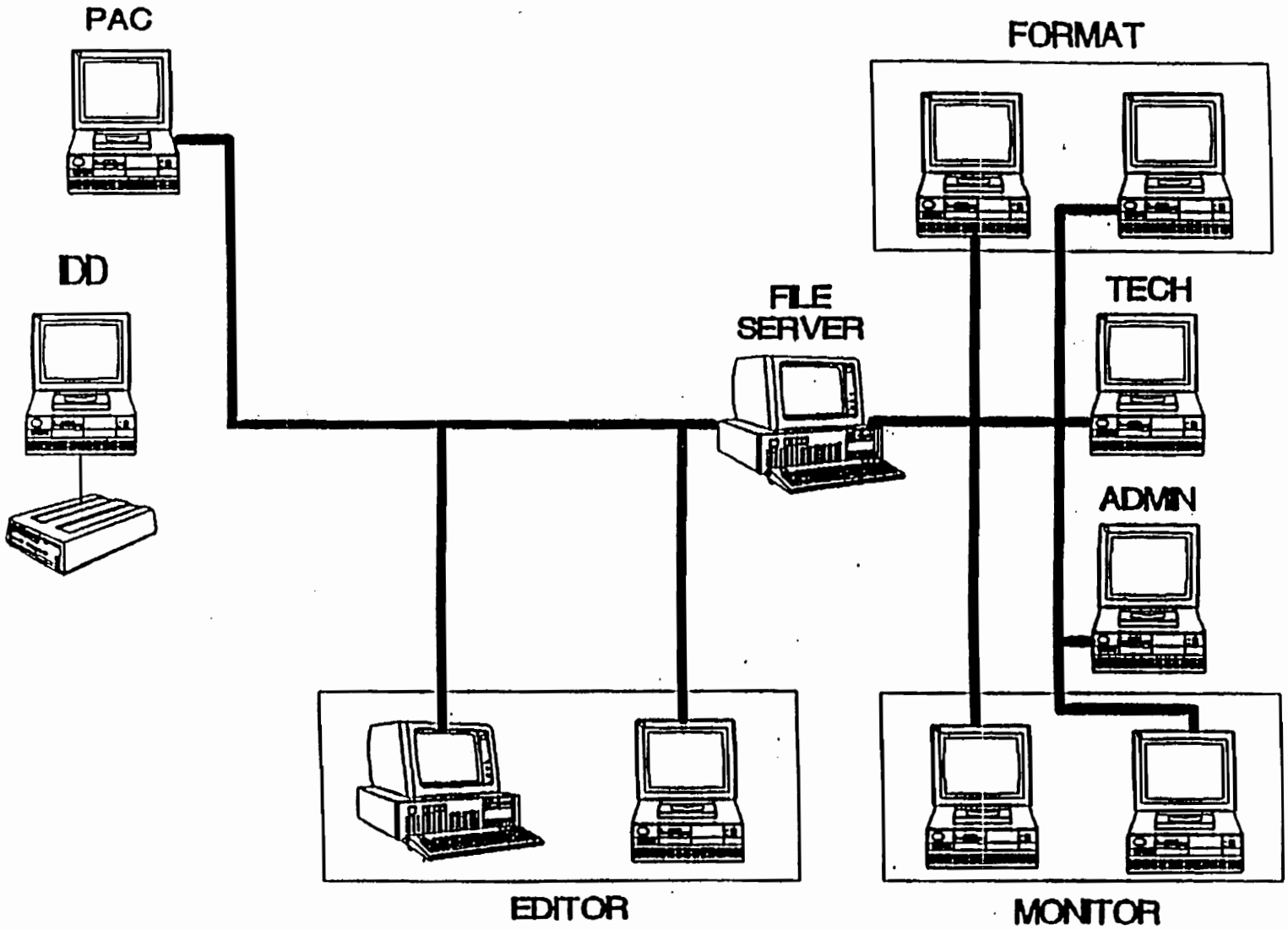

Figure 3. FBIS Bureau Processing System

environment. All text files are in standard ASCII format.

Another FAS developed-program is Clipboard. Clipboard is used by Monitors and Editors to sort and hold incoming press items received by the PAC computer. It allows items to be scanned quickly for important topics, allows quick and easy editing, and automatically stores items onto the LAN server.

1.2.2.2 Monitors, Editors, and Operators

The majority of computers consist of Monitor and Editor workstations. These users coordinate to perform fundamental bureau operations of monitoring, selecting, translating, and editing various items of information the bureau processes.

Items of interest to be processed are selected by Editors and Monitors from the PAC Clipboard, foreign-language media sources such as newspaper articles, or from radio or TV broadcasts. A Monitor translates the article into English and types it into the workstation word processor. The Editor then makes final edits, decides on the priority, and determines one or more destinations for the message. After this, the Editor notifies a Communications Operator that the message is ready for transmission. From the FORMAT workstation, the communication operator adds the necessary header and destination information and sends it to the TRANSMIT machine where it is automatically transmitted to designated consumers.

2.0 DEFINITION OF NEED

Several bureaus process items from English-language hard copy text sources. These sources include newspapers, magazines, facsimile, miscellaneous books and periodicals called "Gray Literature", and translated text from Independent Contractors. Each text article is entered into the FAS processing system by manual means. For example, an Editor selects a newspaper article, photocopies it and gives it to the Communications Operator who retypes it word-for-word into the FAS word processor. FBIS believes that this procedure is too manually-intensive, causes backlogs, and prevents bureaus from meeting their processing requirements. Consequently, FBIS is interested in finding a cost-effective way to make the procedure more efficient.

At this time, FBIS is only interested in evaluating English-language print item processing methods. The translation of foreign material is performed by Monitors and Independent Contractors who are hired to devote full time to those tasks. FBIS plans to reexamine foreign-language processing at a later time in accordance with its strategic plan.

This section describes the problems with manually processing English-language print materials and explains the requirements for a more efficient method. Finally, several alternatives that could meet the needs are evaluated in terms of cost, effectiveness, and technical feasibility. From this analysis, a specific approach is recommended.

2.1 Manual Processing Deficiencies

Seven out of 17 bureaus currently manually process English-language hard copy items. Each bureau has requirements to process a variety of sources and media publications. The amount of material processed, its percent of the bureau's total output, and their sources are shown in Table 1. Also shown is the cumulative expected increase in the amount of material required to be processed over the next 3 years.

As shown in Table 1, FBIS expects the processing requirements to increase an average of 10 % over the next 3 years. The far-east bureaus process the largest amount and expect the largest future increases.

The main job of the Editors and Communication Operators is to select, edit, and prepare items for transmission. As the requirements for processing English-language print material has grown, it has become increasingly difficult to perform

all these tasks effectively, at current personnel and resource levels. A large amount of overtime is spent to meet the processing demands. Nevertheless, the

TABLE 1, Bureau English-language Processing

Bureau Location	Processed English-language Print Items (words per month)	% of total monthly words transmitted per month	English Print Sources	Expected Increase Over 3 year period
Europe	75,000 - 100,000	8 %	English Press Ind. Contractors	10 %
Far-east #1	150,000 - 220,000	25 %	English Press Facsimile Gray Literature	15 %
Far-east #2	100,000 - 150,000	20 %	English Press Ind. Contractors	15 %
Far-east #3	200,000 - 300,000	25 %	English Press Facsimile Gray Literature	20 %
Central America	50,000 - 100,000	15 %	English Press Ind Contractors	10 %
S. Asia	75,000 - 90,000	10 %	English Press	10 %
Middle East	10,000 - 20,000	10 %	English Press	5 %

three far-east bureaus still experience backlogs and delays almost daily in transmitting items. At this time, only the far-east bureau are experiencing these problems.

The amount of time per month spent processing these items is shown in the first column of Table 2. Note that the total processing time consists of Operator typing time plus the time it takes for an Editor to prepare them for processing. The percentage of their time performing these tasks in shown in column two. These numbers are calculated from an average of 600 words per item, an average typing speed of 40 words per minutes, and an average pre-typing preparation time of 10 minutes per item. For example, at 75,000 words per month, the Europe bureau processes 125 item per month for a total preparation time of 21 hours. At 40 words per minute, the typing time is 31 hours. Thus, the total time to process these items per month is 52 hours.

TABLE 2. Required Benefits of New Processing Method

Bureau	Est. Time to Manually Process per Month		% Editor's time	% of Operator's time	Required savings in man-hours (per month)
Europe	52 Mhrs	Typing - 31 hrs- Prep. - 21 hrs	3 %	5 %	Editor: -- Operator: --
Far-east #1	250 Mhrs	Typing - 150 hrs Prep. - 100 hrs	15 %	20 %	Editor: 15 hrs Operator 24 hrs:
Far-east #2	200 Mhrs	Typing - 125 hrs Prep. - 75 hrs	12 %	20 %	Editor: 10 hrs Operator: 18 hrs
Far-east #3	350 Mhrs	Typing - 250 hrs Prep. - 100 hrs	20 %	28 %	Editor: 25 hrs Operator: 35 hrs
Central America	60 Mhrs	Typing - 40 hrs Prep. - 20 hrs	5 %	8 %	Editor: -- Operator: --
S. Asia	50 Mhrs	Typing - 30 hrs Prep. - 20 hrs	< 5 %	< 5 %	Editor: -- Operator: --
Middle East	30 Mhrs	Typing - 20 hrs Prep. - 10 hrs	<5 %	< 5 %	Editor: -- Operator: --

2.2 Operational Needs for New Processing Method

Processing and User Needs

FBIS needs to find a cost-effective approach so that the bureaus can meet their current and future processing requirements. The backlogs, delays, and overtime that is common at the far-east bureaus is unacceptable. The Editors and Communication Operators are required to devote most of their time performing selection, quality control, and formatting tasks, not data entry. As shown in Table 2, as much as 20% of their time is spent on manual processing English-language print items. Based on their ability to perform all their tasks effectively, FBIS estimates that no more than 5% of the Editor's time and no more than 10% of the Operator's time should be spent processing these items. From Table 2, this requirement is currently applicable at the three far-east bureaus. A minimum savings in time is determined from these requirements as shown in Table 2. These values account for the time resource requirements and the expected increase in processing quantity over a 3-year period.

Operational Requirements

Any proposed procedure changes or new technical system to meet these needs must not only save time, but it must also be cost-effective and meet the requirements listed below. Also, any new system or procedure implemented at the far-east bureaus must be fully compatible at other bureaus since they may one day have the same requirement.

- Fully integrated into bureau operational procedures and technical systems
- Compatible with bureau technical systems and text format characteristics
- User-friendly and within bureau personnel skill levels
- Account for FBIS strategic plans and other concurrent or future development efforts

2.3 Evaluation of Alternative Solutions

Four alternatives to improve the manual processing of English-language material are evaluated based on their life-cycle costs and effectiveness in meeting operational and user requirements. The first alternative is the hiring of additional bureau personnel. The second alternative is to make existing procedures more efficient. The third alternative is to receive print material electronically via Press Agencies or through a facsimile-computer interface. The fourth alternative is to use an Optical Character Recognition system to automatically enter text into the computer.

First, the effectiveness and cost parameters of each alternative will be discussed. Next, the relative merits of each alternative will be compared from a common measure and evaluated. From this analysis, one of the alternatives will be selected as the most cost-effective and feasible approach.

Alternative 1: Hiring Additional Personnel

Effectiveness

A simple solution to meet the English-language processing needs is to hire an additional employee. One part-time person working 20 hours a week as an Editor or Operator could provide at least 80 extra man-hours per month for English-language processing. This alternative could be implemented in a short period of time (< 2 months) and would not require any new computer hardware or software.

A person hired for this purpose must alternate between an Editor and a Communication Operator in order to fully support the processing effort. However, this type of position would be unique and would have to be approved under the guidance of FBIS policies. Even if a person is hired from another bureau, training will be required to learn the bureau's English-language processing procedures. Existing Editors and Operators will also have to adjust to the new hire.

A new Editor/Operator requires at least one computer. Space for more computers is not a general concern at the far-east bureaus, but may be an important consideration at other bureaus.

Cost

The cost of hiring one full-time Editor or Operator is approximately $50,000 per year. Given that a part-time employee working 20 hours a week is all that is needed to save the minimum time shown in Table 2, the cost of a part-time person will still cost at least $20,000 per year, not including training or hiring costs. Assuming that bureau space is available, a new computer will be required at the cost of $2,000.

Alternative 2: Changing Existing Processing Procedures

Current manual processing procedures are examined to determine if ways can be found to reduce the amount of handling, or "pre-typing" time. The current English-language processing sequence of steps are as follows:

STEP 1: Editor selects article from hard copy source (e.g., newspaper)

STEP 2: Editor edits item, writes the file name and precedence in the margin

STEP 3: Editor cuts and pastes item onto 8.5" by 11" paper. Item is photocopied and given to the Communication Operator

STEP 4: Operator types item into FORMAT computer

STEP 5: Operator spell checks it, adds header, and hands hard copy back to Editor

STEP 6: Editor makes final edits from a printout or from soft copy

STEP 7: Operator completes header and transmits item

Effectiveness

Steps 5-7 are standard bureau procedures required for all processed items. However, steps 1-4 are specific to English-language print processing. As an alternative, items could be given directly to the Operator, eliminating the steps of cutting, pasting, and photocopying. This would save an estimated 50% of the preparation time, or as much as 50 hours per month at the first far-east bureau. Although this would free up an Editor's time significantly, it would not free up any of the Operator's time. The only way to reduce this time is to increase the Operator's typing skills. Since Operators have already had specific typing training, it is unlikely that any future increase in typing speed can be expected.

The steps of cutting, pasting, and photocopying is done to make the text easier to read for the Operators. For example, a single 8.5" by 11" sheet of paper with various newspaper columns aligned and enlarged is easier to handle and read than simply being given a large bulky newspaper. Bureaus have reported that without the cutting and pasting, text-entry is much more time-consuming and difficult for the Operators.

Cost

This alternative has no significant costs other than an estimated 40 hours of time to document and implement the new procedures. At $25/hour, this cost is $1,000. Since this alternative requires no hardware or software, there is no procurement or maintenance costs.

Alternative 3; Processing via Electronic Means

Since about 75% of the processed English-language print material is from newspapers, it is worthwhile to examine if these items could be received via antenna similar to how many foreign Press Agencies are received. Newspaper items could be stored on the LAN from the PAC computer the same way as other Press Agency items. This could completely eliminate the need for manually entering newspaper text into a computer.

Effectiveness

Unfortunately, most newspapers are received via hard copy because signals are either unavailable or undetectable. Although not currently feasible, this option should remain open and periodically reevaluated. Over time, the media sources may begin sending the articles electronically or a previously inaccessible signal source may become available.

Although facsimile items only account for about 10% of the material processed, it is possible to receive them into a computer via a fax card. This would require a separate computer with a fax card and software to receive and convert the facsimile image to ASCII text. The hardware and software must be IBM-compatible, FAS software compatible, and be integrated with bureau facsimile machines.

Utilizing a facsimile-computer interface capability requires a new computer system, training time, and integration into the Communication section. At 10% of the processing material, the maximum amount of Operator's time that can be saved is 15 hours, using the first far-East bureau as an example.

Cost

Assuming it is feasible, installing new antennas or receiver systems in the hopes of obtaining additional Press Agency broadcasts is a large development effort. The estimated costs are $20,000 to perform modest upgrades or install new antenna/receiver equipment.

Installing a computer with a fax card to automatically process facsimile items costs approximately $1,500 for the card, $2,000 for a computer, and about $5,000 in startup costs (investigation, planning, training, etc.).

Alternative 4: Optical Character Recognition

OCR technology is specifically designed to automate data entry tasks. An OCR system consists of a scanner attached to a computer containing OCR software. The scanner converts a text item into a graphic image. The image is converted into text by OCR software. There are many commercially-available OCR products with a wide range of performance characteristics.

Effectiveness

An OCR system not only eliminates the time require to enter text manually, but it would also reduce the time Editor's spend preparing articles for the

Operators. Most of the cutting, pasting, and photocopying of items would no longer be necessary since the scanner will do all the text entry. Vendor claims of 5 minutes and better to scan about a 500 word text document are common for systems in the $3,000 price range. Estimating a 50% savings in Editor preparation time, an Editor can expect to save approximately 50 hours a month. Estimating 10 minutes to process OCR items compared to an average of 15-20 minutes to process manually, an Operator can expect to save at least 60 hours a month.

At an average of 200,000 words per month (far-east bureaus) and 600 words per item, the number of scans the system must process is approximately 330 a month, or 4000 per year. Scanners in the price range of $1,000 - $2,000 list an average Mean Time Between Failure (MTBF) anywhere between 2000-6000 scans. As a result, the cost of repairing or replacing scanners over the life-cycle and associated inventory requirements will depend on the products chosen.

Furthermore, in order to effectively automate the process, the system must be able to scan various types of printed material (i.e, newspapers, facsimile) with a high degree of accuracy. Based on claims by vendors, many COTS OCR products have accuracy rates above 99% for items such as newspapers. However, the actual savings in time over the manual method must be compared for every media source processed, including the amount of time it saves for both Editors and Operators.

An OCR computer would have to meet operational and technical compatibility requirements. To fit in with current procedures, an OCR system would have to be integrated as part of the Editor/Operator procedures.

COTS products are compatible with the IBM PC platform used at field bureaus. Thus, users should only require training on how to use the scanner, the software, and the new processing procedures.

Cost

The price range of COTS OCR products range from $50,000 for a high volume, high-speed full page scanner to as low as $200 for a hand-held scanner. Neither of these extremes are required. A common flat-bed scanner that automatically scans various paper sizes, claims an accuracy of greater than 99% for newspapers, and can perform thousands of scans before failing has an average cost of $1,700. OCR software costs an average of $800, and an IBM-compatible computer costs about $1,500. Thus, the acquisition cost of an OCR system that meets the minimum requirements is around $4,000. Assuming about $10,000 in startup costs such as

testing and evaluation, and $1,000 a year maintenance costs given MTBF and utilization requirements, the total 3-year life-cycle cost is estimated to be $17,000.

Alternative Comparisons

The discussions of the effectiveness and costs of the four alternatives are summarized in Table 3 for comparison. For this analysis, each effectiveness parameter is given a weighting factor that is applied to all alternatives. The higher the number in the "weight" column of each alternative, the more effective the alternative meets the requirement. Each parameter has a "maximum weight" to show its relative importance compared to other parameters. In addition to effectiveness, a benefit-to-cost ratio for each alternative is computed by dividing the quantified savings in time over the life-cycle cost.

As shown in Table 3, hiring additional personnel is the most effective alternative based on the weighting factor criteria. Although changing bureau procedures or receiving items via electronic means are the least effective alternatives, they are also least-cost. The OCR alternative is very effective in terms of saving time, but its effectiveness is reduced compared to alternative 1 because of the need to integrate and support a new system. Alternatives 3 and 4 require significant considerations of technical compatibility issues while all alternatives require changes to bureau procedures.

The lowest cost alternative is to modify existing procedures. The highest cost alternative is to hire additional personnel. The benefit-to-cost ratio, using $25/man-hour to quantify the savings in time, shows alternative 2 as the highest and alternative 1 as the lowest. Although modifying bureau procedures is low-cost and saves Editor processing time, it does not save any time for the Operator. In fact, it probably increases the time required of an Operator due to the elimination of the Editor cutting and pasting steps. Thus, alternative 2 is deemed not feasible because of the potential reduction in the benefit/cost ratio combined with the fact that it does not meet the minimum requirement to save Operator time.

OCR has a benefit-to-cost ratio of about 6, making it very cost-effective. It also saves the most amount of time, which is the most important factor in evaluating these alternatives. From an overall comparison of costs, benefits, and effectiveness, OCR has the best potential for success. Based on this evaluation, the Optical Character Recognition is recommended as the best approach. The alternative approaches should be reevaluated as needed.

Table 3. Comparison of Alternative Approaches

Evaluation Parameters	Maximum Weight	New Hire (20 hrs/week) Alt #1	Weight	Modify Manual Procedures Alt #2	Weight	Electronic Processing (fax card) Alt #3	Weight	Optical Character Recognition Alt #4	Weight
EFFECTIVENESS									
Save Editor Time	30	40hrs/month	25	50 hrs/month	27	0 hrs	0	50 hrs/month	27
Save Operator Time	30	40 hrs/month	22	0 hrs	0	15 hrs/month	10	75 hrs/month	29
Software Compatible	20	none required	20	none required	20	FAS and ASCII compatible	10	must be compatible with FAS	10
Hardware Compatible	20	New computer	20	none required	20	IBM-compatible	20	Must be IBM compatible	20
Maintainable/Supportable	20	Hire Editor or Operator only	7	No maint. costs	10	Hardware/software support	12	inventory and vendor support needed	12
Personnel Skills/Training	10	Train new hire	7	Train users	9	New system training	5	within skills levels	5
User-Friendliness	10	no change	10	More difficult data-entry	3	facsimile-computer interface	5	scanner-software interface	5
Procedure Impact	10	Editors & Operators	8	New Procedures	5	change facsimile procedures	5	New procedures	5
LIFE-COST (3 years)		$62,000		40 man-hours ($1,000)		$8,500		$17,000	
Benefit/Cost Ratio ($25/man-hour saved)		Benefit/Cost = 24,000/62,000		Benefit/Cost = 45,000/1,000		Benefit/Cost = 13,500/8,500		Benefit/Cost = 112,500/17,000	
		Effectiveness: 119 Benefit/Cost = .26		Effectiveness: 95 Benefit/Cost = 45		Effectiveness: 67 Benefit/Cost = .4		Effectiveness: 113 Benefit/Cost = 6.6	

3.0 OCR SYSTEM TECHNICAL AND USER REQUIREMENTS

An OCR system approach is recommended based on a preliminary study of alternative solutions. Implementing this approach requires that the exact technical requirements of OCR system components be described and quantified. The requirements described in this section are intended to provide input for evaluating different COTS OCR products and for establishing design criteria.

The requirements for an OCR system for the field must encompass all necessary technical, performance, and maintenance functions. However, it is also important to consider bureau operations, to understand how an OCR system must operate and coexist with existing bureau systems. Figure 4 illustrates how OCR system requirements are broken down and presented in this section. Based on user requirements and a preliminary operations scenario, OCR technical, and performance requirements are established. These requirements are intended to be used for evaluating and selecting OCR preliminary designs for field testing. After testing proposed OCR systems under operational conditions, the requirements should be reviewed again and changed if necessary so that they are consistent with a final system design configuration.

3.1 OCR Technical Requirements

An OCR system consists of a scanner, a computer, and processing software. The technical requirements of this system are a function of how it will be used and how items must be processed and integrated into the FAS computer system. First, the operational scenario is described of how an OCR system must operate in the FAS environment. Then, the hardware and software functions required of the OCR system elements are described.

3.1.1 Hardware/Software Interfaces

Operations Scenario

The integration of an OCR computer system requires the correct technical and operational interactions with the FAS processing system. Although the process of selecting text items would not significantly change, Communication Operators would instead use a dedicated OCR computer system to automatically scan printed items into the word processor, instead on manually typing them in. Functionally similar to a MONITOR computer, new items would be input and stored on the LAN through the OCR system and then processed like other messages. Since

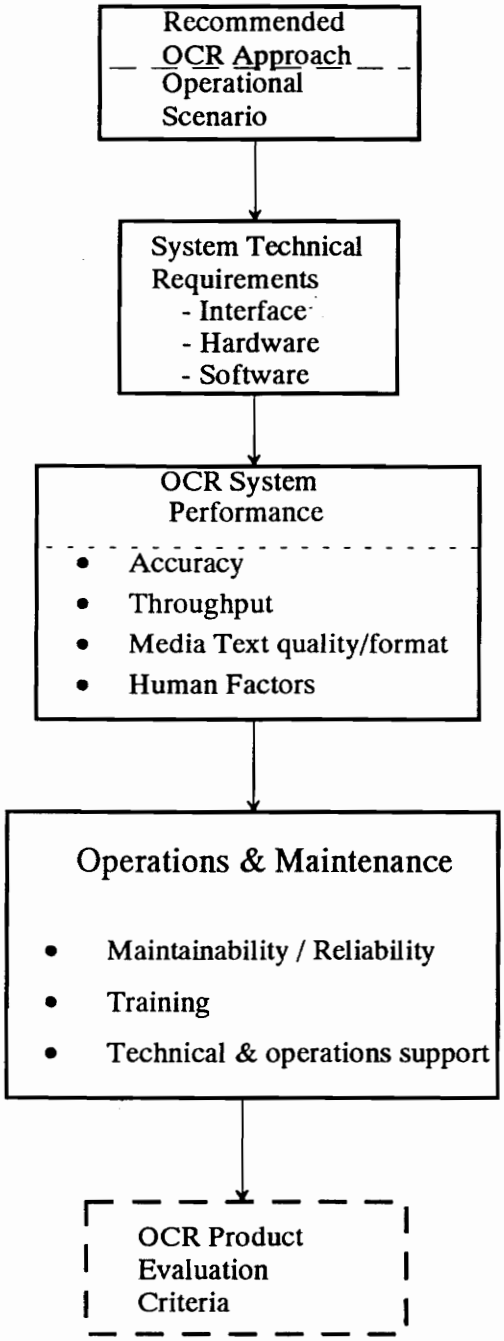

Figure 4. OCR Requirements Breakdown

all items are processed using the FAS software, the OCR systems must be designed within its constraints.

OCR computer system operations will be performed by bureau Communications Operators. The required user and processing procedures for the OCR system are outlined in Figure 5. OCR items are scanned and stored on the OCR computer and transferred to the LAN where it is processed, formatted, and transmitted like other messages to designated users. Items to be scanned are simply marked by the Editors and handed to the Operators for entry into the OCR computer. The operator places the item on the OCR scanner. The item is processed by the OCR software, converted to ASCII text format, given a file name, and stored on the LAN.

The FAS computer must interface with the LAN so that messages can be stored and processed properly. Figure 5 illustrates one such design in which items are sent to the PAC computer through the OCR computer's serial port. This way, items can be processed like routine Press Agency items. A serial cable is connected from the OCR computer into a device called a STARGATE which contains 8 ports for connection to Press Agencies. In standalone mode of operation, scanned items must be stored on floppy diskettes and manually copied onto the FORMAT computer.

Compatibility

To integrate the OCR system into the processing environment as discussed, the following interface requirements must be met:

- OCR hardware and software must be fully compatible with FAS hardware and software
- OCR system must be compatible with FAS operating systems, text formats, and Press Agency port characteristic
- OCR processing procedures must be in accordance with existing FAS and bureau operational procedures
- Fully compatible with proposed future processing system hardware and software

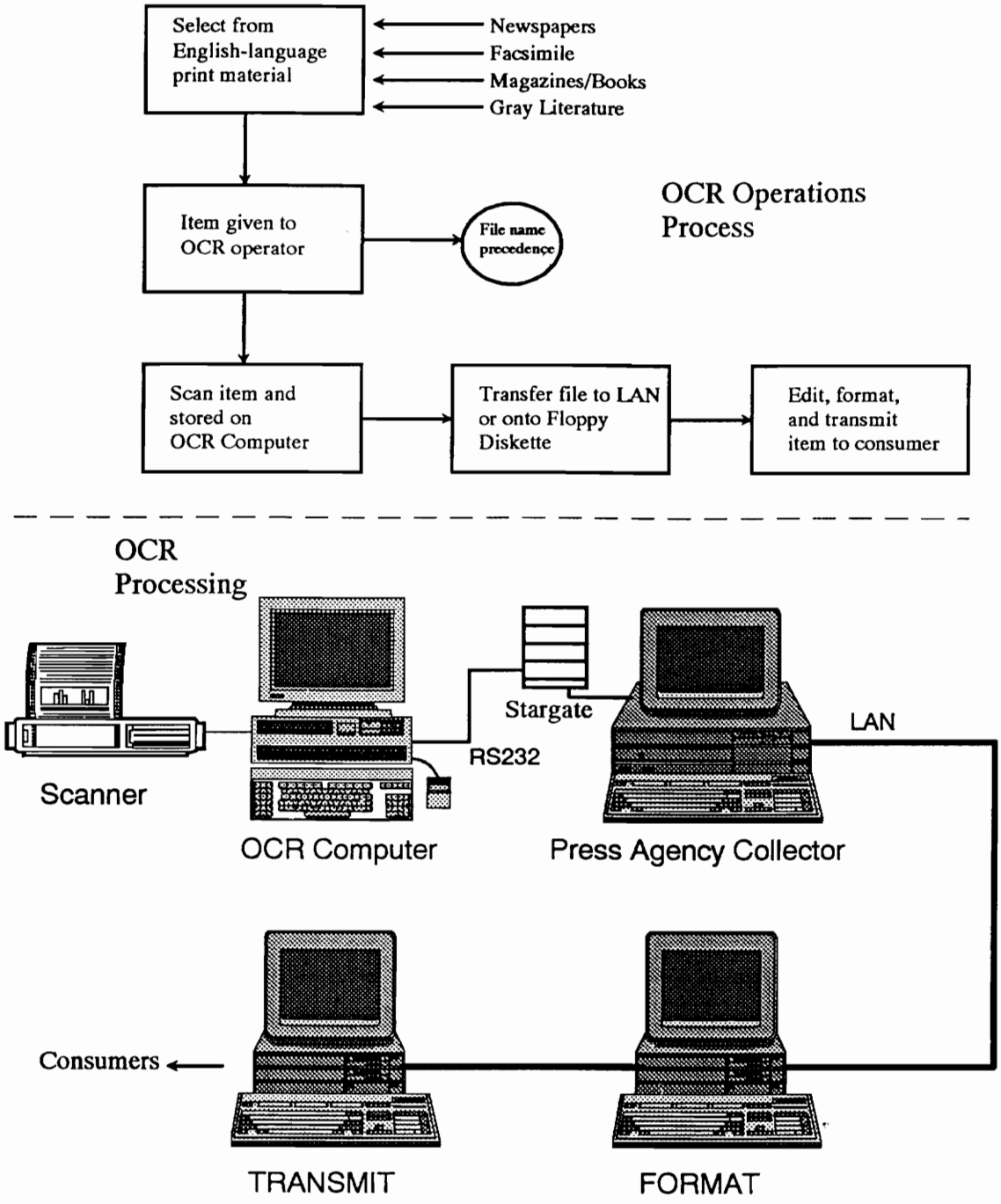

Figure 5. OCR Operational Scenario

3.1.2 Hardware/Software Requirements

To perform the operational tasks and be integrated into the FAS processing system, the OCR hardware and software must meet specific technical requirements. The characteristics of OCR system components must also consider the required different text sources to be scanned.

OCR Scanner

With the wide variety of English-language print media sources received by bureaus, the OCR scanner must be very flexible and rigorous in its ability to scan various fonts, pitches, paper quality, and formats. As a result, the OCR scanner must also meet the following minimum input specifications:

- accept from 5" by 5" up to 8.5" by 14" copy, hold at least 20 pages, and have an automatic feeder that does not require user intervention
- scan papers of various weights (newspapers to books) without jamming
- allow at least 256 shades of grayscale (color is not required) and provide contrast and brightness control selection (scanning graphics not required)
- flatbed scanning (ease of use)
- continuously adjustable sheet-size settings

OCR Computer

To maintain standardization and flexibility with existing computer systems and to account for the expected use and FAS interface requirements, the OCR computer must meet the following requirements:

- IBM-compatible 386 or 486-class computer using same specifications, where applicable, as existing FAS computers supported in the field
- contain enough RAM to maximize performance of the OCR system
- provide a sufficient hard disk drive to store OCR computer software and at least one week's worth of scanned text items (required for standalone and as a backup).

- provide at least VGA-compatible, 16-inch, .28 dot pitch monitor with a resolution of at least 1024 X 768 (allows easier on-screen processing of small font text)

- provide RS232 serial port to allow the ability to interface with the PAC computer and provide at least two expansion slots for upgrades and connectivity to printers and/or other communications devices

- provide 5.25" and 3.5" floppy disk drives compatible with OCR software

OCR Software

All processing of scanned items is done by OCR software residing on the OCR computer. The functional requirements of this software must account not only for how items must be processed but also its compatibility with the FAS software. It is also important to account for the large variety of media source characteristics. As a result, the OCR software must have the following characteristics:

- automatically convert all items scanned into the system into the ASCII text format used by the FAS word processor (XyWrite)

- process all required text and media sources such as newspapers and facsimile and meet specified media, font, and character discrimination requirements.

- provide the following statistics on demand:
  - accuracy rate (errors per 1000 characters or words)
  - conversion time (words or characters per minute)
  - known errors

- allow automatic spell checking with user-definable dictionary

- deferred processing (to allow control over processing start time)

Given the variety of media sources, it is important to tailor OCR parameters in order to optimize it use. With newspapers and facsimile being the two major hard copy sources that traditionally have the poorest quality text, the software must allow the user to change, store, and load multiple software configurations that optimize the processing of these items. The OCR software must have the following selectable settings available which can be loaded before scanning:

- page format (multiple or single page)
- manual/automatic column settings (for newspapers)
- select ASCII output file type (to be compatible with FAS word processor)
- choose output file name with at least 11 characters in length with extension (every item must have a file name before being formatted and transmitted)

3.2 OCR Effectiveness and Performance

The performance of an OCR system for field bureau depends on the variety of text sources to be scanned and changing text parameters such as fonts, pitch, and character quality. Effectiveness must not only be measured quantitatively in terms of accuracy, but also from the user's perspective of handling and ease of use. Given the time required to manually process items and the desired time savings expected from an OCR system, a required accuracy rate must be determined for the various media sources.

3.2.1 Accuracy/Throughput

The throughput of an OCR system is defined as the number of accurately recognized characters per second. Although this depends on the number of errors and accuracy rate, it depends primarily on the quantity and size of the documents to be scanned. With a maximum of 20 text items to be scanned in one day (given about 200,000 words per month and an average of 600 hundred words per item), and an required maximum utilization of 7 hours per day, the scanner must take no more than 20 minutes to scan a 600 word item.

Even if a document has excellent quality text, errors will still occasionally occur in the finished document. Although manufacturers claim an accuracy rate of scanned characters above 99%, it is often based on the best quality text. Since items scanned by the bureau are not always high quality by any means, accuracy rates will vary.

Items to be scanned by the bureaus seldom exceed 1000 words. At a total of about 5000 characters for a 1000 word document, the number of errors for an accuracy rate of 99% is 50. Although spell checking could correct some errors automatically, most will require correction by scanning the item line-by-line with the original document.

Estimating 5 minutes to correct 10 errors, as much as 20 minutes could be spent correcting errors even with an accuracy of 99%. Although the actual time needed

to correct errors must be established during field testing, it is required that the OCR system be designed to exceed a 99% accuracy rate so that the total processing time is less than the manual processing time as required in Table 2. To record accuracy information, the system must provide the following statistics upon request:

- % errors per 1000 characters or words
- % substitution and unidentified characters per 1000 characters or words

3.2.2 Print Discrimination/Ease of Use

All English-language printed items to be scanned are black and white text. The capability to scan graphics or color documents is not required. The materials to be scanned have different potential quality problems that require different OCR system characteristics.

Newspapers

There are several potential problems with scanning newspapers. First, since newspapers are generally very thin and double-sided, there is a significant possibility of character bleed-through which could cause character errors. Secondly, since OCR uses background to character grayscale differences to identify characters, streaks and smears commonly associated with newspapers may cause problems. Not only must the scanner be able to handle these quality problems with minimum errors, but methods (such as photocopying before scanning) that reduce errors must be considered. Lastly, newspaper articles often run on separate pages and in a column format, which could require more time to prepare items for scanning. Because of these characteristics, the OCR system must provide the following features:

- column selection and formatting (newspaper column format)
- on-screen text selection or text blocking
- OCR products explicitly specify their support of newspaper scans

Facsimile

The quality of facsimile material is significantly poorer than the original copy. The quality of the received item depends on the original text and on the

performance of both fax machines. Problems with skewing, incomplete characters, and light text are common problems. Although manual methods such as photocopying should be attempted to improve text quality before scanning, the effectiveness of OCR for these items must be established during testing.

Consequently, the OCR hardware and software must be compatible with personal computer facsimile hardware and software products. This will allow the potential alternative of automating the input of fax items into the OCR computer via a fax card, bypassing the OCR system. For facsimile scanning, however, the OCR system must be at least state its ability to scan facsimile material with an accuracy above 99%.

Books/magazines/Publications

These sources share many of the problems that come with newspapers. In addition, they include many varieties of paper sizes, fonts, pitches, and colors. The cutting and pasting required to put together magazine articles for scanning, for example, requires that the scanner be very user-friendly in how different paper is placed on the scanner and how sections of text are selected.

With the wide variety of fonts and styles associated with these types of material, the OCR system must support a font range of 6-28 pitch.

Media Discrimination

Since accuracy is also a function of the quality of the text, not all English-language sources may not be of sufficient quality for use in the OCR system. Through field testing, a determination must be made as to which media sources are effective for regular OCR use. Based on these results, the operational requirements may be re-evaluated or other alternatives analyzed. The testing and evaluation criteria for determining OCR system effectiveness in media discrimination is further explained in sections 4 and 5.

3.3 OCR Maintenance/Support Requirements

In addition to technical requirements, maintenance and support requirements must be specified to ensure that the system is maintainable and useful over the life-cycle of the system. FBIS and bureau personnel resource requirements to support this effort must also be identified.

3.3.1 Reliability and Life-span

Deployment and Life-Cycle

The hardware purchased for the OCR system should meet an expected 3-year life-cycle. The system is expected to be deployed 2-3 months after completion of field testing and evaluation. Bureau operational hours at the far-east bureaus can extend to around the clock. Therefore, the OCR system must be able to function 24 hours a day. The maximum amount of downtime, scheduled or unscheduled, must be less than 2 hours a month.

From Table 1, we see an average of 200,000 words keyed in by the far-east bureaus each month. With 600 words per article on the average, we have,

$$\text{Ave. \# of scanned items per day} \quad \frac{\text{200,000 words per month}}{\text{22 work days*600 words per item}} = 15$$

It is required that the OCR scanner be useful over a minimum of 12 months before requiring replacement or repair. At an average of 15 scans per day, or 330 a month, the mean time between failure (MTBF) of the scanner must be at least 4000 scans. A one-year warranty is required for the scanner and computer.

In case of failure, the required workaround is to go back to the current manual method of data entry. A spare scanner is required on-site as a backup. No additional computer spares are required since it is compatible with existing bureau computers. Any failed equipment will be shipped for repair with a total turnaround time not to exceed one month. However, with daily use and the complexity of the system, access to technical support from the vendors is required during the initial installation and testing of the system. Given the expensive phone costs overseas, technical support by phone will not be feasible for an extended period of time. Therefore, the system must be easy enough to install without extensive vendor support.

3.3.2 Training/O&M

Prior to installation, the Communication Operators require training on how to use the new system. One person at the bureau is required to be responsible for training users. Since this responsibility requires technical expertise, the trainer should be the bureau Technician, Engineer, or Systems Administrator. Preferably,

this person will also function as the project leader and be responsible for coordinating the installation, transition, and testing tasks.

The time required of bureau personnel to support the OCR investigation effort is estimated to be one man-month. The test bureau personnel must state their availability before testing can start. The estimated time required of an FBIS person to coordinate and perform systems engineering tasks is estimated to be one man-month.

Before implementing the OCR system, the skill levels of the users must be identified. The design of the system must not be too complicated or too much of a hindrance for the OCR Operators. Even though the Operators are familiar with computers, the current functions are simple menu-driven commands accessed via keyboard. In contrast, the OCR system may involve a Windows Operating System environment, use of a mouse, modifying processing parameters, optimizing software performance, and using error correction processing methods. The operators should not be required to program software or change hardware and cables as part of normal operations.

Editors as well as the OCR Operators must transition to the new method of operations. Editors, as part of the selection process, will be an integral part of the OCR English-language press selection and processing environment. The installation of a new OCR system must not require additional time on their part or a significant change in their procedures.

The level of effort involved in training must be planned and organized according to the following general training requirements:

- Editor and OCR Operator training on new OCR operational procedures
- OCR Operator training on the use of OCR scanner and software, including training on selecting pre-programmed system settings for maximizing performance and accuracy
- Pre-processing and post-processing training such as formatting and spell checking
- Media discrimination and selection training
- OCR program and utility programs training
- FAS system interface training (LAN and standalone)

Based on the design proposal and training requirements, a formal training plan must be written. In addition to the above requirements, it must include a checklist and instructions of how to use the system. The time and resources necessary to implement the training plan must also be specified.

4.0 EVALUATION METHODOLOGY FOR OCR PRODUCTS

The first step in determining a feasible OCR design is to choose COTS OCR products based on their ability to meet the stated requirements. To select products, some kind of evaluation system for the OCR performance effectiveness must be established so that the ability of a product to meet the requirements is measurable. When selecting OCR products, the acquisition and life-cycle cost of the product must also be determined and factored into account. This section describes a method whereby the effectiveness and costs for different OCR products can be compared from a common measure.

After selecting OCR products based on this criteria, the evaluation of a proposed system design proceeds to the field testing phase. After field testing, the product evaluation criteria can again be used to make changes to the preliminary design. The result should be a final OCR system design that is applicable over specified print items and identified bureaus.

4.1 Evaluation Criteria

Weighting factors will be applied to evaluate and select OCR products. The process is similar to the way the four alternative approaches were evaluated in section 2.3. Each stated hardware, software, performance, and maintenance requirement will be given a numerical weight corresponding to its relative importance on system effectiveness. The more important the requirement, the higher the weight. Non-tangible requirements such as training and skill levels should also be weighted but since they generally depend on the results of a demonstration or from field testing, they should be addressed after a proposed OCR system is field tested.

The second criteria used to choose an OCR system design is life-cycle cost. The required acquisition and maintenance costs of various selected OCR system components over its life-cycle are determined for each proposed design.
The summary of weighted factors for effectiveness will be compared to the total life-cycle cost of each OCR design alternative so that the most feasible OCR design is selected.

4.1.1 System Effectiveness/Weighting Factors

A scale of 1 to 10 is used to weigh the relative importance of OCR effectiveness factors. A weighted factor of 10 is used to designate the highest priority, or "must have" requirements. Values less than 10 correspond to those

parameters that are deemed non-critical and considered of lesser importance compared to others.

The effectiveness parameters are organized into hardware, software, accuracy, and reliability/maintainability sections. Each of these sections has an additional multiplying factor in order to weigh the relative importance of each of the OCR components. For example, the OCR scanner is given a multiplying factor of 2 while the OCR computer is given a factor of 1. Each effectiveness parameter listed under an item is multiplied by the multiplying factor and summarized to obtain a final weighted effectiveness number for the OCR system. The higher the final sum, the more effective that particular OCR system alternative.

Most of the effectiveness parameters are a stated feature of the product. However, with the unique bureau operations and the large variety of media sources and styles, several support and performance requirements such as accuracy and human factors depend to a large extent on the results of field testing. Therefore, when initially comparing products, the relative dependence on field testing is estimated by giving each effectiveness factor a Field Weighting Factor number. The higher the number, the higher its dependence on field testing. Although subjective, this factor can be used to choose between two products that are very close in their comparative merits.

Table 4 illustrates how these weighted factors are applied to the functional and maintenance effectiveness characteristics required for an OCR system. Note the multiplication weight factors applied to the hardware, software, accuracy, and reliability sections, and the Field Weighting Factor.

4.1.2 Life-Cycle Cost Analysis

Each cost parameter associated with an OCR system under investigation must be added up so an effective tradeoff analysis can be achieved with the effectiveness factors weighed in Table 4. The life-cycle cost parameters of an OCR system are summarized in Figure 6. The systems engineering costs are assumed to be the same for every alternative evaluated (one man-month). Acquisition, installation, training, and maintenance costs are added up in dollars where $25/hour is used as the fixed cost per man-hour. Although these factors depend on the system design, it is assumed that each of the costs have the same relative weight so that the final costs are simply added up into one sum.

Table 4. OCR Effectiveness Weighted Factors

OCR System Requirement	Weighting Factor (WF)	Field Test Factor (FTF)	OCR System Requirement	Weighting Factor (WF)	Field Test Factor (FTF)
HARDWARE			SOFTWARE (weight=3)		
Scanner(weight=2)			OCR Software		
- Flatbed	8	n/a	- compatible with scanner &computer hw/sw	10	n/a
- Text/monochrome	10	n/a	- process Newspaper and facsimile (stated accuracy >99%)	10	10
- automatic feeder	6	5	- automatic spell checking	5	n/a
- 5'by5' to legal size	8	n/a	w/user-defined dictionary	7	
- contrast/brightness control selection	10	3 2	- statistics - conversion rate	7	n/a
- 100-300dpi	8	n/a	- # errors	8	3
- 256 shades of gray	7	2	- selectable/programmable settings	10	3
- 6-28 font pitch	9	5	page format	8	2
Reliability/Maintainability (wt=1			columns	8	n/a
- MTBF of 7900	7	n/a	output file name	9	n/a
- 24-hour operation	7	n/a	- ASCII conversion	10	n/a
- <2 hr/month maint.	3	n/a	- on-screen text selection	7	2
- 1 yr warranty	5	n/a	- deferred processing	5	n/a
			Operating System (weight=1)		
OCR Computer (wt=1)			- compatible with OCR software and hardware	10	n/a
- IBM/FAS-compatible 386/486	7/10	n/a	Accuracy/Throughput (wt.=4)		
- VGA 1204x768 monitor	5	n/a	- >99% accurate for all source items	10	10
- RS232 port	10	n/a	- <10 minutes per scan (1000 words)	8	8
- High-density drives	7	n/a			
- High storage drive (>100Meg)	4	3			
Summary	WF:	FTF:	Summary	WF:	FTF:

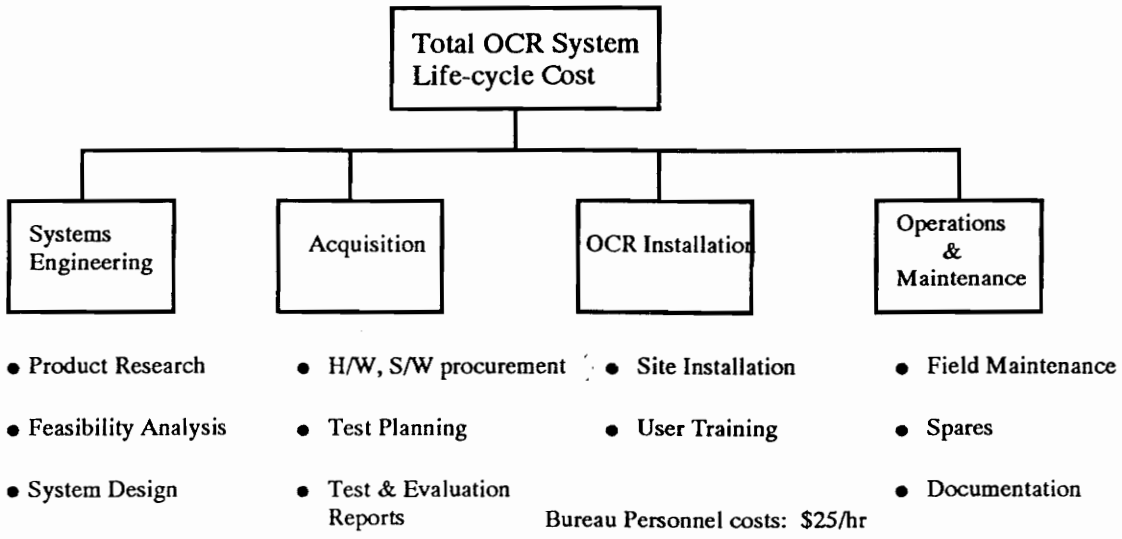

Figure 6. OCR Life-Cycle Cost Parameters

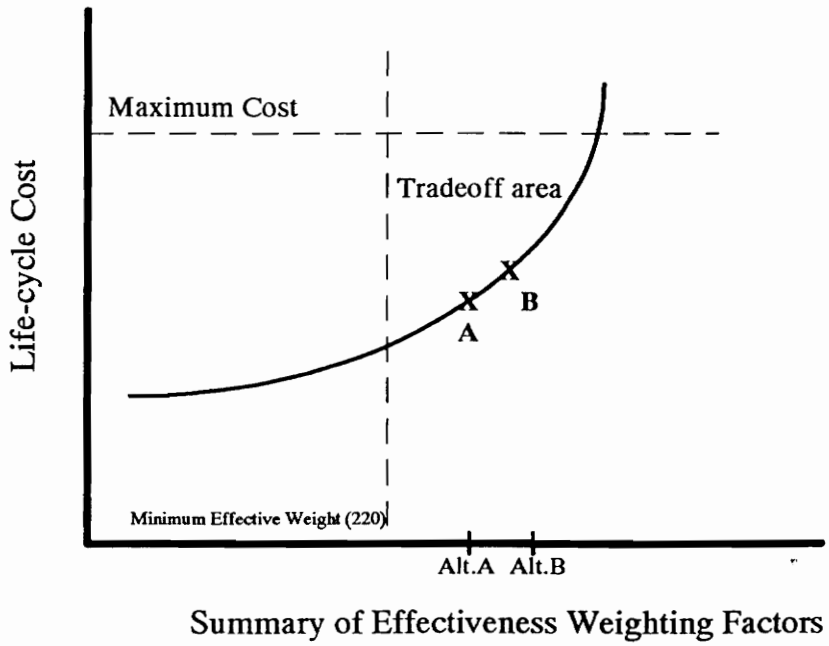

Figure 7. Weighted Factors/Cost Evaluation Criteria

Both the summary of weighting factors and life-cycle cost for each alternative can be compared as shown in Figure 7. The area under the budget limitation and within the minimum effective weight is the tradeoff area. The minimum effective weight is the sum of all weighted factors of 10 listed in Table 4. In general, the highest weight within budget is the optimal choice.

4.2 Evaluation Technique

The Appendix gives an example of how this effectiveness and life-cycle cost criteria is applied. Two OCR systems, comprised of various COTS products, are compared. Their relative costs and effectiveness will be plotted and evaluated using Figure 7. At least two systems should be selected using this approach so they can be tested under field conditions.

5.0 FIELD TEST PLAN

The second step in determining a feasible OCR system design is to test a proposed OCR system design at a field bureau, where it can be operated and evaluated under real operational use conditions. Field testing in a realistic bureau environment is crucial to determine the feasibility of an OCR system design and to identify limiting and optimizing factors. Test plans and procedures are required to ensure that all necessary data and test scenarios are planned. Field testing helps determine the optimal OCR design, which media sources can and cannot be processed through the system, and helps defines what further actions may be required before implementing a final design.

Test Scenario and Schedule

It is recommended that at least two alternative OCR scanners and OCR software packages be tested during a three month period. The same tests should be performed on each system. After completion of the test, a test report should be submitted which summarizes the results. This on-site, or "Type II" testing should be conducted at one of the far-east bureaus.

Test Personnel

Test personnel shall include individuals who operate and maintain the system along with supporting bureau technical staff. It is the responsibility of the project leader to ensure that all necessary test data is collected and documented. Testing begins after the initial off-line training is complete and the hardware and software is installed. When operations begin, there must be constant feedback and interaction between the project lead and the users so that the data is collected regularly and accurately.

5.1 Performance Tests

To determine the effectiveness of the OCR system, tests must be performed on all the various media sources, text qualities, fonts, and formats received by the bureau. All performance tests must be performed on each of the following types of items:

- Newspapers
  - single and multiple page articles
  - articles with multiple columns

- articles with different fonts, styles, and text quality

- Facsimile
  - sample from each source
  - facsimile vs. photocopy
- Magazines, books, and periodicals
  - various fonts, paper type, format
  - minimum and maximum paper size and colors

It may be difficult to quantify and determine specific characteristics such as font size and text style. If this is the case, as many samples of different media types as possible should be tested so that the maximum range of the material processed by the bureau is tested.

The following performance test data should be obtained for the media sources identified:

- Accuracy: % unidentified characters per 1000 words
  % substituted characters per 1000 words
  % errors per 1000 characters
  - Data must be collected daily over the test period so that any increase in errors due to scanner degradation can be noted.

- OCR overhead time (i.e. scan load time) for minimum and maximum item size and type
  - Quantify in number of minutes per item

- comparisons between pre-processing, post-processing, and total processing time - quantify in # minutes per item

- OCR throughput
  quantify in number of accurately recognized characters per second

- number of errors per media type and words per item

- number of post-processing errors from using OCR processing (e.g., spell check) versus using the FAS word processor

- increased number of errors when text is slanted, skewed, or rotated

- quantity of errors from OCR software ASCII translation

- range of media types, sizes, and styles over which OCR system is effective - list time saved for each item compared to manual entry

5.2 Optimization Tests

During testing, the user selected scanner and software settings are tested to obtain the optimal settings for the various media types and sources. The objectives of these tests is to determine how to optimize the performance of the OCR system. The collection of performance data must include the following tests for optimization:

- time to scan and accuracy for the various OCR scanner contrast, brightness, and dots per inch (dpi) settings for the various media sources
- accuracy improvement from photocopying facsimile, newspaper, etc.
- improvements from using different paper or facsimile machines
- compared times to process when editing using OCR software versus FAS software
- quantify any other operator actions that reduces errors and increase OCR performance
- % of media that can be processed unattended

Any bureau-developed software used to optimize the processing should be included in the bureau's test report. The bureau must submit a Request for Change (RFC) for any proposed changes involving interfacing with the processing system LAN.

5.3 Reliability and Human Factors Tests

Not only must the OCR system be accurate and fast, but it must also be manageable, reliable, and easy to use. It must be user-friendly in terms of installation, training, and maintenance. The following information is required during the test period:

- describe and compare skill levels required of users
- amount of training required and relative user-friendliness of system

- COTS vendor product installation and troubleshooting support availability and cost
- number of software or hardware failures and corrective action taken
- any degradation of scanner quality over test period
- amount of human intervention to transfer items to FAS system

5.4 Preliminary Test Procedures

Test procedures are required to plan and implement the testing requirements. The test bureau should provide test procedures that describe how the test data will be collected. Figure 8 illustrates the test and evaluation approach. All performance and optimization data should be recorded daily in notebooks or a spreadsheet.

Any problems or corrective actions taken should also be documented so that they can be evaluated along with the test data for any necessary procedure or design changes. During the test period, the bureau should submit regular status reports of the testing progress.

5.5 Criteria for Success

After completion of the testing phase, a report summarizing the testing results will be used to measure the success of the OCR system. The overall criteria for success is the system's effectiveness in processing various English-language print material based on its required savings in time over manual methods. The test data may reveal that certain media sources cannot be processed by the OCR system effectively. For these items, the alternatives outlined in section 2 should be reexamined as needed.

The test report should also report any technical constraints or limiting factors inherent in the design. To help in the analysis, the bureau should submit hard copy samples of the different printed items used in the testing. At least one sample of each media type and its results will help in evaluating problem areas and in finding ways to increase the accuracy of the OCR process.

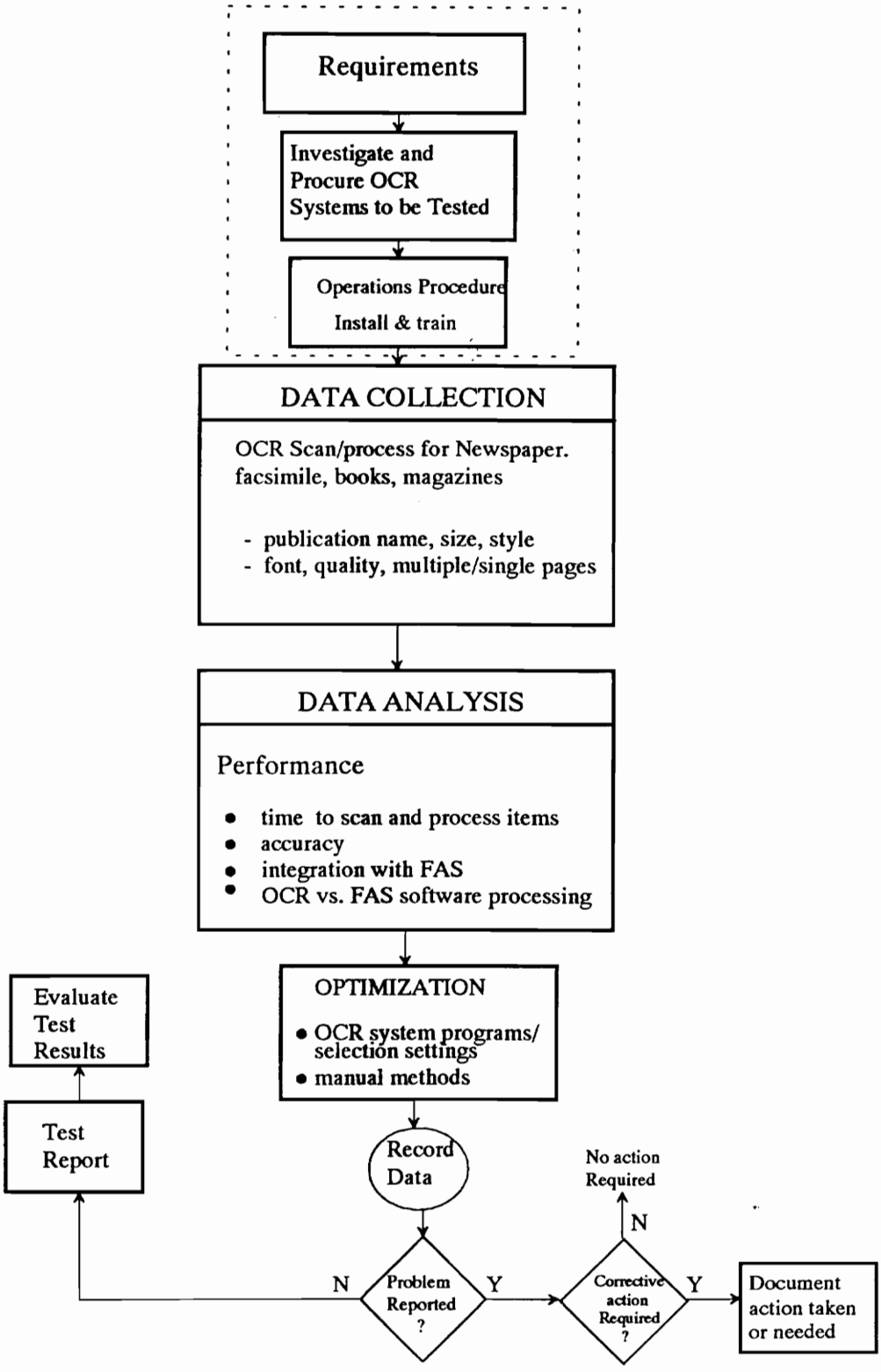

Figure 8. Field Test & Evaluation Procedures

6.0 CONCLUSIONS AND RECOMMENDATIONS

This report studied the requirements for a new way to process English-language print materials at FBIS field bureaus. Based on an evaluation of alternatives and the selection of an OCR design approach, this report presented methods for selecting feasible COTS OCR products and for determining its specific applicability and effectiveness at field bureaus. A review of OCR technical requirements, OCR product evaluation criteria, and a field test plan was presented in support of these objectives.

The Appendix presents the results of preliminary research into COTS OCR hardware and software products. Two OCR designs are compared to show how cost and weighting factor criteria are applied to select the most cost-effective OCR products. In addition, it is recommended that before procuring products for field testing, that the vendor demonstrate its performance claims, ease of use, skills required, and relative ease to install and maintain.

After field testing and receiving the test report, all necessary information should be available to determine the following:

- relative OCR usefulness over the various print items, types and styles, quantity of words in an article, and quality of publications
- the specific circumstances where a proposed OCR system is cost-effective and applicable at other bureaus
- whether alternative solutions need further investigation for those publications that cannot be processed effectively by OCR
- management, operations, and technical constraints

Figure 9 illustrates a schedule designed to meet the FBIS deadline of 1 January 1994 for implementing a new English-language processing system at FBIS bureaus. The schedule shows the key steps involved in implementing the recommended OCR design, including milestones such as the Preliminary Design Review (PDR) and the Critical Design Review (CDR), where applicable.

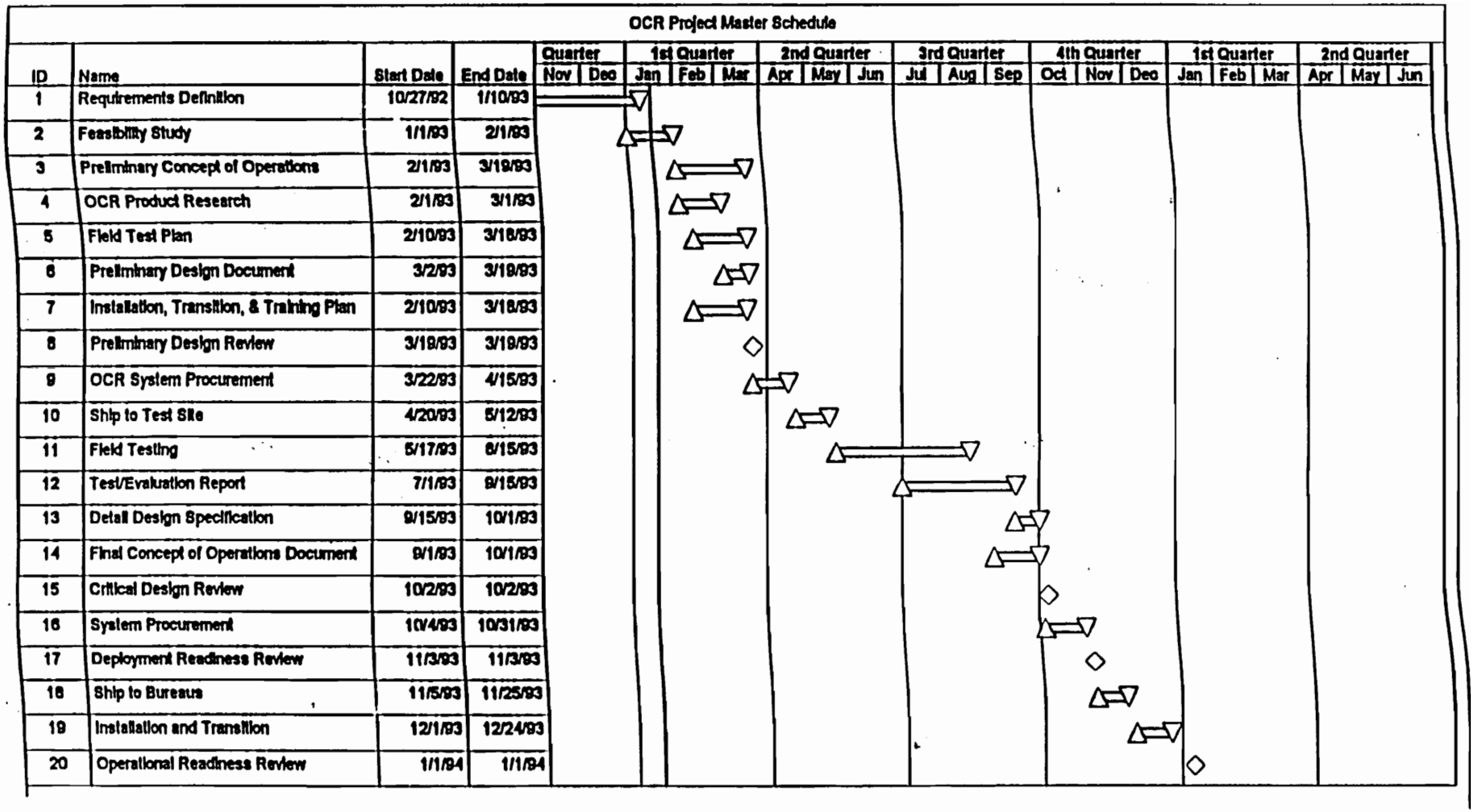

Figure 9. OCR Project Schedule

APPENDIX: OCR Product Information

The product information summarized in Tables 5 and 6 is used to show an example of how COTS OCR products are evaluated and selected based on the criteria presented in section 4. The features of five OCR scanners and four OCR software packages are summarized in Tables 5 and 6 along with their procurement costs.

TABLE 5. OCR Scanners

Requirement	ScanJet Plus	MicroTek	Abaton 300	Datacopy	AVR 3000
	$1,400	$1,100	$1,800	$2,300	$2,500
Functions					
Scan Resolution (dpi)	75,150,300	3-300	72-300	75-300	75-300
Gray-scale modes	16 or 256	256	16 or 256	256	16 or 256
Brightness, Contrast Control	255,255	14,14	256,256	15,5	128,128
Performance					
Time to scan one page of text @300dpi	19 seconds	>300 seconds	200 seconds	37 seconds	100 seconds
Mean # of scans before failure	5000	4,000	N/A	7,000	N/A
Max Scan Size	8.5 by 14	8.5 by 14	8.5 by 14	8.5 by 14	8.5 by 14
Automatic Document Feeder	Yes ($595)	Yes ($624)	Yes ($595)	Yes ($595)	Yes ($795)
Bin Capacity (#sheets)	20	50	50	50	100
Continuously Adjustable Sheet-size settings	No	n/a	Yes	Yes	Yes
Minimum Sheet-size	7.7 by 10.1	n/a	4 by 4	5.5 by 5.5	5 by 5
Scan Buffer size (default, max)	32K, 32K	64K, 64K	32K, 32K	28K, 28K	512K, 512K

n/a = data not availble/pending

Table 6. OCR Software Characteristics

Feature	Omni-Page Professional	TypeReader	ExpressReader	WordScan
	$995	$695	$595	$495
Column formatting	Yes	Yes	Yes	Yes
Output format				
- ASCII text conversion	Yes	Yes	Yes	Yes
- Microsoft Word	Yes	Yes	Yes	Yes
Font range (pitch)	N/A	5-64	6-64	6-60
User-definable templates	Yes	Yes	N/A	N/A
Retain Text Styles	Yes	Yes	Yes	Yes
On-screen text selection	Yes	Yes	Yes	Yes
Automatic Spell Checker	Yes	Yes (user-defined dictionary)	Yes	Yes
Deferred Processing	Yes	N/A	Yes	Yes
Automatic Learning		Yes	N/A	N/A
Statistics	Yes	No	Yes	N/A
Windows compatible	Yes	Yes	Yes	Yes
Scanner compatibility				
HP ScanJet Plus	Yes	Yes	Yes	Yes
Mictotek	Yes	Yes	Yes	Yes
Abaton 300	Yes	Yes	Yes	Yes
Datacopy GS Plus	Yes	Yes	Yes	Yes
Xerox	Yes	No	Yes	Yes
Process text received via fax card in computer	Yes	Yes	N/A	N/A

N/A = data not available

OCR System Evaluation Example

The following two systems will be evaluated based on the weight and cost criteria outlined in section 4:

SYSTEM ALTERNATIVE A:

Computer: DTK 486 with the following specifications:

- 8 MB RAM
- NEC 16" VGA .28 pitch monitor
- 120Mbyte hard disk drive
- 101 Enhanced keyboard with mouse
- DOS 5.0 and Windows 3.1
- 2 parallel, one serial port,
- 8 expansion slots
- high-density 5.25" and 3.5" floppy drives

Operating System: MS Windows 3.1 & DOS 5.0

OCR scanner: Hewlett Packard ScanJet Plus (Monochrome)

OCR Software: TypeReader

SYSTEM ALTERNATIVE B:

OCR Computer: EPSON 386/25Mhz with the following specifications:

- 40 MByte hard disk drive
- 14" VGA .28 pitch POWER II monitor
- 4MB RAM
- two parallel, one serial port
- 8 expansion slots
- high-density floppy drives
- 101 keyboard w/mouse

Operating System: MS Windows 3.1 & DOS 5.0

OCR Scanner: DataCopy GS Plus
OCR Software: OmniPage Professional

LIFE-CYCLE COSTS:

Procurement Cost

The two systems have the following procurement costs:

System Alternative	System component	Cost
A	DTK 486	$2,700
A	ScanJet Plus	$2,185 (with automatic feeder)
A	TypeReader	$ 795
		$5,680
B	EPSON 386	$1,700
B	DataCopy GS Plus	$2,285 (with automatic feeder)
B	OmniPage Professional	$ 995
		$4,980

Training Cost

The OCR software requires the most training time due to its complexity. Since OmniPage is generally more complex than Type Reader, it is estimated that 40 man-hours are needed for alternative A and 60 hours are needed for alternative B. At $25/hour, this gives $1,000 for alternative A and $1,600 for Alternative B.

Installation Cost

Although functionally different, it is assumed that a total of 15 man-hours is required to install both systems for a total of $375.

Spares/Maintenance

Over a three-year life-cycle with 4000 scans on average per year, the mean number of repairs depends on the MTBF of the scanner and computer. For Alternative A, a MTBF of 3500 scans means a scanner failure can be expected about every 12 months. This gives a total of 3 failures over a three year period. Assuming $1,000 per action, the repair cost is $3,000. Data concerning the time

required for routine maintenance is unavailable, but for simplicity it is assumed to be the same for both alternatives.

For alternative B, the scanner MTBF is 5,000 which gives an average of two repairs for a cost of $2,000. The maintenance costs of the computers for both alternatives is assumed to be $500 over the 3-year period.

Testing/Documentation

It is assumed that the same amount of time is required to test and document the operations of both systems. It is assumed that two bureau personnel spend 20% of their time on these tasks over a 6 week period. This gives total of 48 hours for a cost of $1,200.

Startup Costs

The costs of performing startup tasks such as investigating products and writing procedures is estimated to be one man-month, or $10,000.

Total Life-Cycle Cost

Adding these costs, the total life-cycle cost for each alternative is as follows:

ALTERNATIVE A:	Acquisition	- $5,680
	Training	- $1,000
	Installation	- $ 375
	Spares/Maintenance	- $3,500
	Testing/Documentation	- $1,200
	Systems Engineering	- $10,000
	TOTAL:	$21,755

ALTERNATIVE B:	Acquisition	- $4,980
	Training	- $1,600
	Installation	- $ 375
	Spares/Maintenance	- $3,500
	Testing/Documentation	- $1,200
	Systems Engineering	- $10,000
	TOTAL:	$20,655

Effectiveness Factors

Using the information in Tables 5, and 6, the total weighted factors for each of alternatives A and B are shown in Table 7 and Table 8. The total weight is a function of the OCR system ability to meet the stated requirements.

The life-cycle cost and weighted factor data is plotted in Figure 10. As shown, alternative A has the most cost and least weight while Alternative B has least cost and most weight. Consequently, alternative B would be selected over alternative A. From Table 7 and Table 8, the Field Weighting Factor of Alternative B is higher than alternative A.

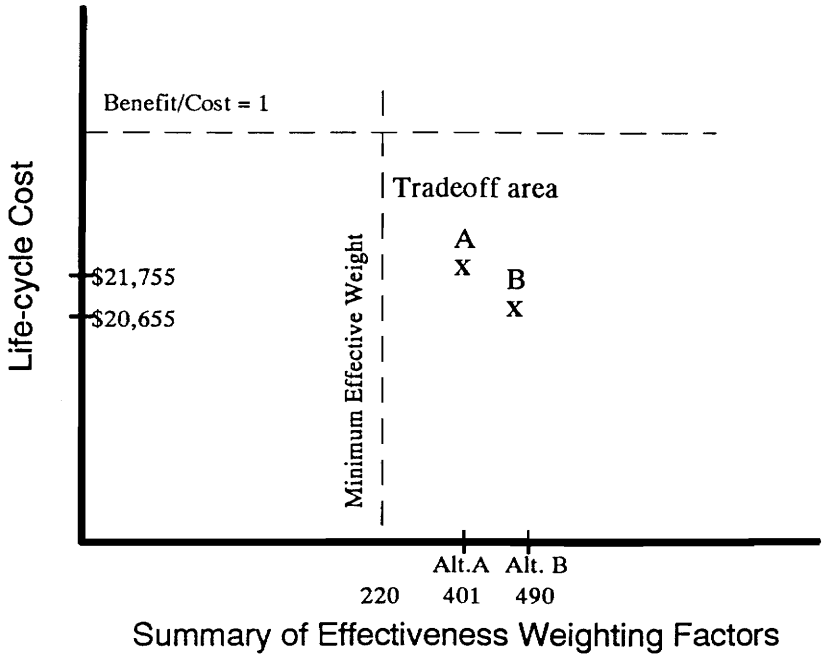

Figure 10. Alternatives Effectiveness/Cost Chart

Table 7. OCR Effectiveness Weighted Factors for Alternative A

OCR System Requirement	Weighting Factor (WF)	Field Test Factor (FTF)	OCR System Requirement	Weighting Factor (WF)	Field Test Factor (FTF)
HARDWARE			SOFTWARE (weight=3)		
Scanner(weight=2)			OCR Software		
- Flatbed	8 ✔	n/a	- compatible with scanner &computer hw/sw	10 ✔	n/a
- Text/monochrome	10 ✔	n/a	- process Newspaper and facsimile (stated accuracy >99%)	10 ✔	10 ✔
- automatic feeder	6 ✔	5 ✔	- automatic spell checking	5 ✔	n/a
- 5'by5' to legal size	8	n/a	w/user-defined dictionary	7 ✔	
- contrast/brightness control selection ✔	10 ✔	3 ✔ 2	- statistics - conversion rate	7	n/a
- 100-300dpi	8 ✔	n/a	- # errors	8	3
- 256 shades of gray	7 ✔	2	- selectable/programmable settings	10 ✔	3
- 6-28 font pitch	9	5	page format	8	2
Reliability/Maintainability (wt=1			columns	8	n/a
- MTBF of 7900	7	n/a	output file name	9	n/a
- 24-hour operation	7 ✔	n/a	- ASCII conversion	10 ✔	n/a
- <2 hr/month maint.	3	n/a	- on-screen text selection	7 ✔	2
- 1 yr warranty	5 ✔	n/a	- deferred processing	5	n/a
OCR Computer (wt=1)			Operating System (weight=1)		
- IBM/FAS-compatible 386/486	8/10 ✔	n/a	- compatible with OCR software and hardware	10 ✔	n/a
- VGA 1204x768 monitor	5 ✔	n/a	Accuracy/Throughput (wt.=4)		
- RS232 port	10 ✔	n/a	- >99% accurate for all source items	10 ✔	10
- High-density drives	7 ✔	n/a	- <10 minutes per scan (1000 words)	8 ✔	8 ✔
- High storage drive (>100Meg)	4 ✔	3 ✔			
Summary	WF: 146	FTF: 19	Summary	WF: 255	FTF: 62

TOTAL WF: 401
TOTAL FTF: 81

Table 8. OCR Effectiveness Weighted Factors for Alternative B

OCR System Requirement	Weighting Factor (WF)	Field Test Factor (FTF)	OCR System Requirement	Weighting Factor (WF)	Field Test Factor (FTF)
HARDWARE			SOFTWARE (weight=3)		
Scanner(weight=2)			OCR Software		
- Flatbed	8 ✔	n/a	- compatible with scanner &computer hw/sw	10 ✔	n/a
- Text/monochrome	10 ✔	n/a	- process Newspaper and facsimile (stated accuracy >99%)	10 ✔	10 ✔
- automatic feeder	6 ✔	5 ✔	- automatic spell checking	5 ✔	n/a
- 5'by5' to legal size	8 ✔	n/a	w/user-defined dictionary	7	
- contrast/brightness control selection	10 ✔	3 2	- statistics - conversion rate	7 ✔	n/a
- 100-300dpi	8 ✔	n/a	- # errors	8 ✔	3
- 256 shades of gray	7 ✔	2 ✔	- selectable/programmable settings	10 ✔	3
- 6-28 font pitch	9	5	page format	8	2
Reliability/Maintainability (wt=1			columns	8 ✔	n/a
- MTBF of 7900	7	n/a	output file name	9 ✔	n/a
- 24-hour operation	7 ✔	n/a	- ASCII conversion	10 ✔	n/a
- <2 hr/month maint.	3	n/a	- on-screen text selection	7	2
- 1 yr warranty	5 ✔	n/a	- deferred processing	5 ✔	n/a
			Operating System (weight=1)		
OCR Computer (wt=1)			- compatible with OCR software and hardware	10 ✔	n/a
- IBM/FAS-compatible 386/486	8/10✔	n/a	Accuracy/Throughput (wt.=4)		
- VGA 1204x768 monitor	5 ✔	n/a	- >99% accurate for all source items	10 ✔	10 ✔
- RS232 port	10 ✔	n/a	- <10 minutes per scan (1000 words)	8 ✔	8 ✔
- High-density drives	7 ✔	n/a			
- High storage drive (>100Meg)	4	3			
Summary	WF: 162	FTF: 17	Summary	WF: 328	FTF: 102

TOTAL WF: 490
TOTAL FTF: 119

REFERENCES

"A Cursory Look at Optical Character Readers & FBIS Headquarters Requirements", Briefing Notes, 12/6/91.

Far-East Bureau - OCR Test & Evaluation Report, December 1992.

FBIS Publications List, Processing Requirements, 17 July 1990.

"The Very Best in OCR", Imaging Magazine, volume 1, March 1992, pgs 43-47.

"OCR Products Evaluation", PC Magazine, August 1991, pgs 306-360.

Computer Shopper, September 1992.

FAX Buyer's Guide, Bedford Communication, Inc. NY, 1992.

"A Historical Review of OCR Research & Development", Proceedings of the IEEE, July 1992.