Creating and Validating a Large Image Database for METTREC

Michael D. Garris and William W. Klein

National Institute of Standards and Technology

ABSTRACT

The National Institute of Standards and Technology (NIST) is in the process of setting up a new series of conferences named the Metadata Text Retrieval Conferences (METTREC). It will focus on evaluating document conversion using optical character recognition (OCR), and information retrieval (IR) technologies. Evaluations will be designed to investigate the impact of machine recognition errors upon information retrieval and to determine what interfaces are appropriate to integrate the two technologies. To implement this conference, we require databases that can be used for conference evaluations and has chosen the *Federal Register* to be the initial document source. It is a large, complete set of documents containing metadata that will allow quantitative evaluation of recognition and retrieval technologies. This paper describes the activities associated with scanning the *Federal Register* and validating the document images within the database. The process of image validation includes translating filenames, assuring image integrity, and verifying correct page sequences. In order to reduce the cost of validation, we minimized human resource expenditure by exploiting OCR and high-speed visual adjudication from images by an operator. This process minimizes the expensive handling of paper to validate document image collections.

Keywords: CD ROM, document, image database, information retrieval, METTREC, optical character recognition, OCR, quality, scanning, technology evaluation

1.0 INTRODUCTION

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) has conducted extensive research in optical character recognition [1] and text retrieval technologies [2][3][4]. In both areas, a number of conferences have been held to evaluate and understand the state of the technology [5][6][7].

In keeping with prior evaluation methods, the ITL under joint sponsorship with the Department of Defense is in the process of setting up a new series of conferences named the Metadata Text Retrieval Conference (METTREC). It will focus on evaluating document conversion using optical character recognition (OCR) and information retrieval (IR) technologies within the context of integrated tasks. Evaluations will be designed to investigate the impact of machine recognition technology upon information retrieval and to determine what interfaces are appropriate to integrate the two technologies.

This effort requires a database that can be used for conference evaluations. The *Federal Register* for 1994 was chosen to be the source for this database because it is: (1) a complete set of documents within the public domain; (2) a large collection containing over 250 issues consisting of over 67,000 pages of information; (3) a structured document set whose hierarchy contains metadata; (4) a collection of pages containing significant variations in print and image quality; and (5) a set of documents for which the text for the entire collection is stored within electronic files. Although the latter is not part of this paper, the text stored within electronic files will allow us to derive the "ground truth"; this represents the correct

ASCII characters that an OCR and IR system should recognize and retrieve and will allow us to quantify recognition and retrieval accuracy.

To conduct evaluation conferences, involving scientific experiments, training and testing materials must be rigorously prepared. These materials are comprised of two types of data: (1) document images and (2) document text and metadata tags. This paper focuses on scanning and validating document images. A validated image must represent the correct paper page, contain an untruncated bitmap, have reasonable pixel dimensions, and contain an acceptable range of rotational skew. With a large document collection, this is a tedious and expensive process. OCR and other automated techniques can be used to minimize the cost of preparing these materials. This paper documents the process by which we scanned and validated the *Federal Register* image database that will be used for METTREC.

2.0 FORMAT OF FEDERAL REGISTER

The United States Government Printing Office (GPO) prints the *Federal Register* each work day to record the transactions of the government. The *Federal Register* is the official daily publication for Rules, Proposed Rules, and Notices of Federal agencies and organizations, as well as Executive Orders and other Presidential Documents. Each issue is printed and bound into a book(s). Usually, the GPO publishes one book per day that is printed in mostly 9 point Vermilion font. During 1994 due to the printing volume, the GPO printed multiple books on April 25 and November 14. Each book contains three distinct sections:

- **Prefix:** Prefix pages consist of three types of pages: a hard cover page, a soft cover page, and content pages. Figures 1 through 3 provide illustrations of prefix pages. Figure 1 illustrates a hard cover page that contains the date, the volume number, an address label area, and a postal class identification; it is printed on high grade kraft paper to minimize or prevent damage incurred by postal processing and handling. As illustrated in Figure 2, a soft cover page identifies the date, the volume number, and the page numbering sequence for the body pages contained within the book. A content type of page, illustrated in Figure 3, contains a page heading that includes a page number which is instantiated as an upper case Roman numeral. The first content page contains general information with regard to the *Federal Register* and its usage. Each subsequent content page identifies the contents of the *Federal Register* book for the day. Both soft cover and content pages are printed on recycled newspaper quality of paper.

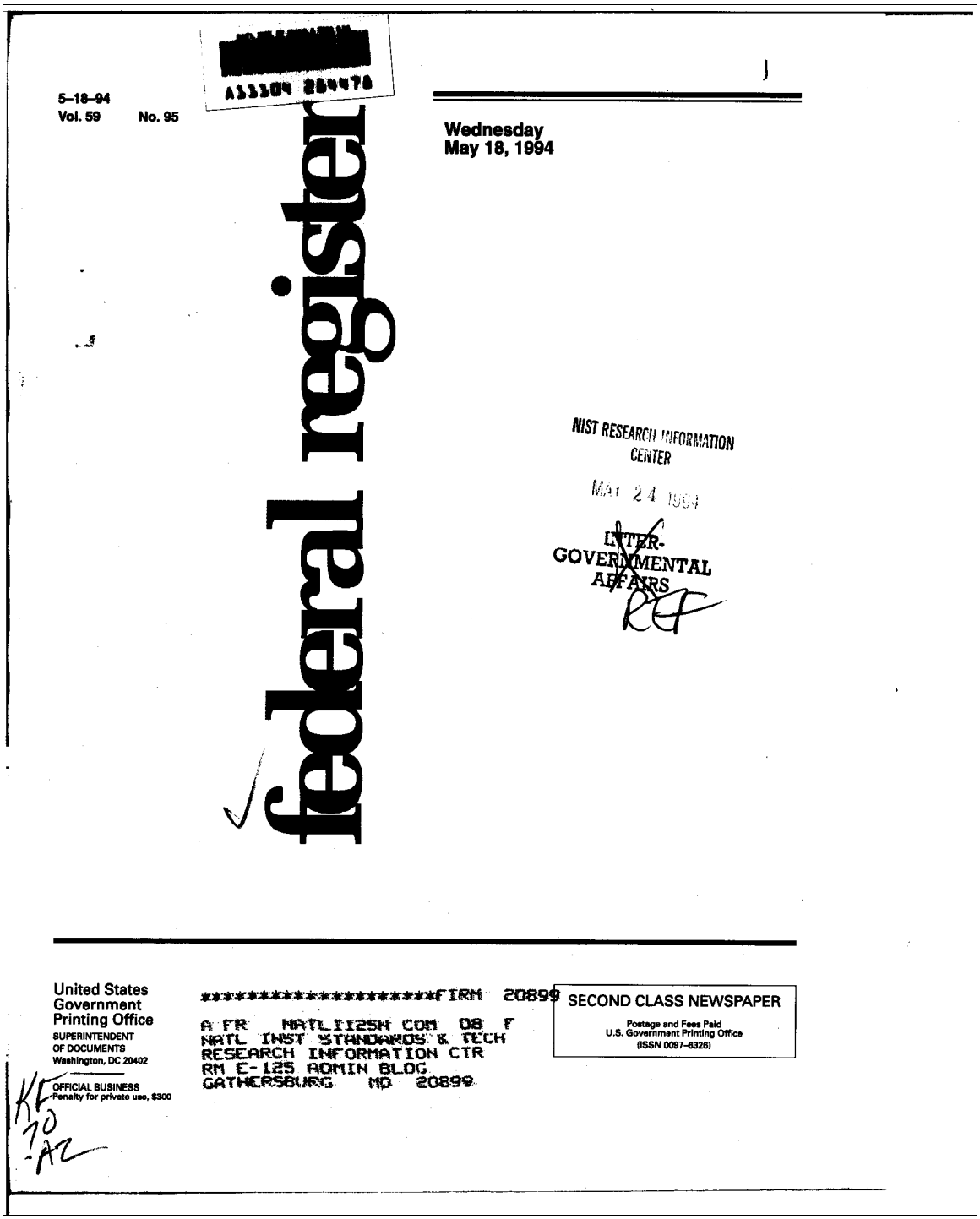
- **Body:** Figures 4-5 illustrate typical body pages contained within a book. A body page provides a record of the meeting notices, proposals, and transactions of the United States government for the day. There are two type of body pages: section and detail. A section type of page, illustrated in Figure 4, is similar in appearance to a soft cover page and is used to divide the *Federal Register* into distinct sections by topic. It contains the name of the issuing agency, the Code of Federal Record (CFR) title and part(s) affected, and a brief description of the specific section subject; this type of page does not contain a page number field. A detail page, illustrated in Figure 5, elaborates Presidential and Executive Order(s), Rules and Regulations, Proposed Rules, and Sunshine Act Meeting Notices. Each page contains a page heading that includes a page number which is instantiated as an Arabic number. Both section and detail pages are printed on recycled newspaper quality of paper.

- **Appendix:** Figure 6 illustrates a typical appendix page contained within a book. The appendix consists of pages that provide reader aids which allow a reader of the *Federal Register* to access information and to index specific information contained within multiple previously published *Federal Register* book(s). An appendix page contains a page heading that includes a page number which is

instantiated as an lower case Roman numeral. Appendix pages are printed on recycled newspaper quality paper.

With the exception of cover and section pages, each page of the *Federal Register* is printed with a page heading that includes a text banner printed above two horizontal lines. The text banner line contains information that identifies the document, the volume, the date, the topic, and a page number.

Each *Federal Register* book ends with a blank hard cover page that minimizes or prevents damage incurred by postal processing and handling.



5-18-94
Vol. 59 No. 95

federal register

Wednesday
May 18, 1994

NIST RESEARCH INFORMATION CENTER
MAY 24 1994
INTER-GOVERNMENTAL AFFAIRS

United States Government Printing Office
SUPERINTENDENT OF DOCUMENTS
Washington, DC 20402

OFFICIAL BUSINESS
Penalty for private use, $300

FIRM 20899
A FR NATLI25N COM 08 F
NATL INST STANDARDS & TECH
RESEARCH INFORMATION CTR
RM E-125 ADMIN BLDG
GAITHERSBURG MD 20899

SECOND CLASS NEWSPAPER
Postage and Fees Paid
U.S. Government Printing Office
(ISSN 0097-6326)

Figure 1. Typical Hard Cover Page.

Figure 2. Typical Soft Cover Page.

FEDERAL REGISTER Published daily, Monday through Friday, (not published on Saturdays, Sundays, or on official holidays), by the Office of the Federal Register, National Archives and Records Administration, Washington, DC 20408, under the Federal Register Act (49 Stat. 500, as amended; 44 U.S.C. Ch. 15) and the regulations of the Administrative Committee of the Federal Register (1 CFR Ch. I). Distribution is made only by the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.

The **Federal Register** provides a uniform system for making available to the public regulations and legal notices issued by Federal agencies. These include Presidential proclamations and Executive Orders and Federal agency documents having general applicability and legal effect, documents required to be published by act of Congress and other Federal agency documents of public interest. Documents are on file for public inspection in the Office of the Federal Register the day before they are published, unless earlier filing is requested by the issuing agency.

The seal of the National Archives and Records Administration authenticates this issue of the **Federal Register** as the official serial publication established under the Federal Register Act. 44 U.S.C. 1507 provides that the contents of the **Federal Register** shall be judicially noticed.

The **Federal Register** is published in paper and 24x microfiche format. The annual subscription price for the **Federal Register** paper edition is $444, or $490 for a combined **Federal Register**, Federal Register Index and List of CFR Sections Affected (LSA) subscription; the microfiche edition of the **Federal Register** including the Federal Register Index and LSA is $403. Six month subscriptions are available for one-half the annual rate. The charge for individual copies in paper form is $6.00 for each issue, or $6.00 for each group of pages as actually bound; or $1.50 for each issue in microfiche form. All prices include regular domestic postage and handling. International customers please add 25% for foreign handling. Remit check or money order, made payable to the Superintendent of Documents, or charge to your GPO Deposit Account, VISA or MasterCard. Mail to: New Orders, Superintendent of Documents, P.O. Box 371954, Pittsburgh, PA 15250–7954.

There are no restrictions on the republication of material appearing in the **Federal Register**.

How To Cite This Publication: Use the volume number and the page number. Example: 59 FR 12345.

SUBSCRIPTIONS AND COPIES

PUBLIC	
Subscriptions:	
Paper or fiche	202–783–3238
Assistance with public subscriptions	512–2303
Single copies/back copies:	
Paper or fiche	783–3238
Assistance with public single copies	512–2457
FEDERAL AGENCIES	
Subscriptions:	
Paper or fiche	523–5243
Assistance with Federal agency subscriptions	523–5243

For other telephone numbers, see the Reader Aids section at the end of this issue.

"THE FEDERAL REGISTER

WHAT IT IS AND HOW TO USE IT

FOR: Any person who uses the Federal Register and Code of Federal Regulations.

WHO: The Office of the Federal Register.

WHAT: Free public briefings (approximately 3 hours) to present:
1. The regulatory process, with a focus on the Federal Register system and the public's role in the development of regulations.
2. The relationship between the Federal Register and Code of Federal Regulations.
3. The important elements of typical Federal Register documents.
4. An introduction to the finding aids of the FR/CFR system.

WHY: To provide the public with access to information necessary to research Federal agency regulations which directly affect them. There will be no discussion of specific agency regulations.

CHICAGO, IL

WHEN: June 9 at 9:00 am
WHERE: Ralph Metcalfe Federal Building
Conference Room 328
77 West Jackson Blvd.
Chicago, IL
RESERVATIONS: 1–800–366–2998

Figure 3. Typical Content Page.

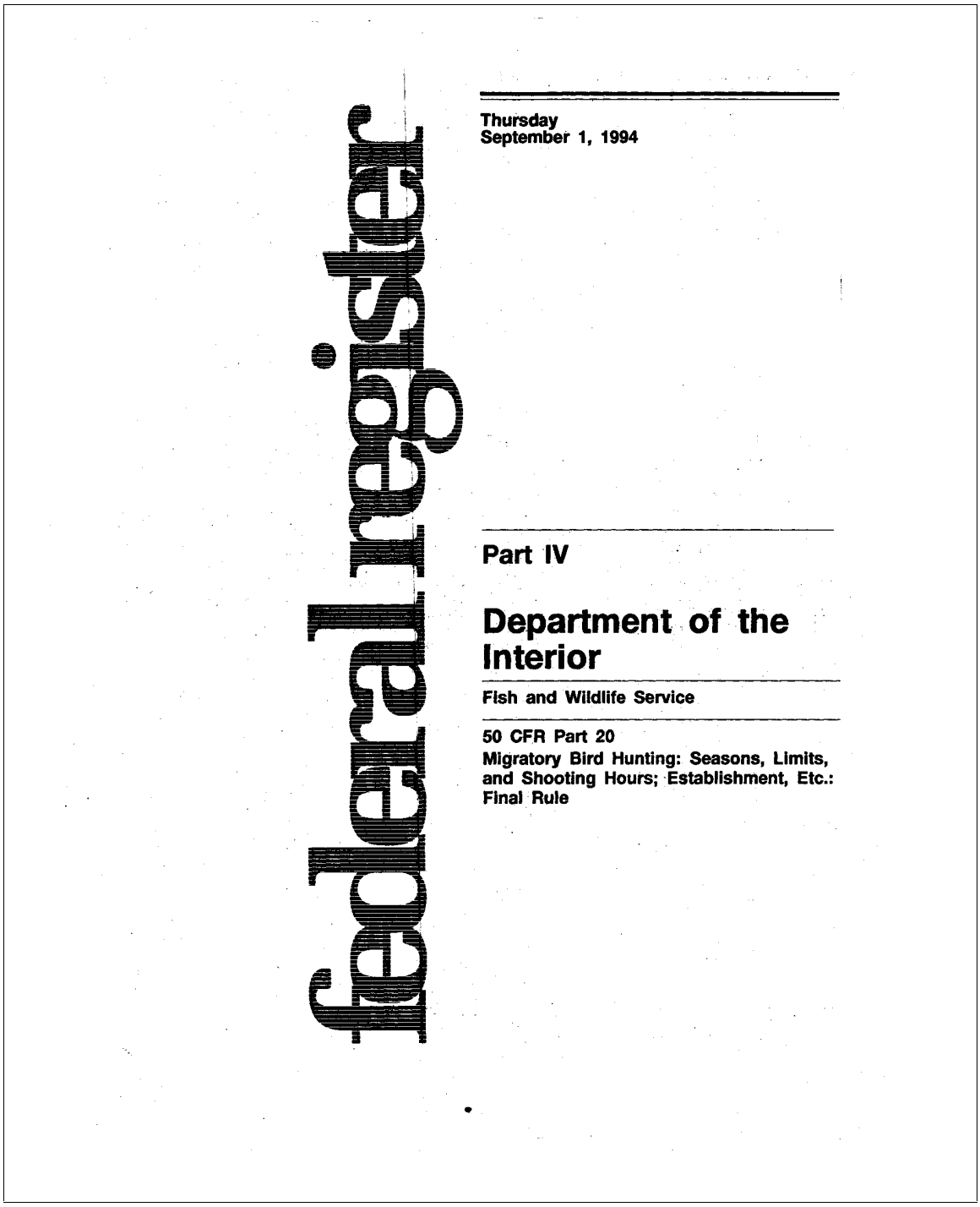


Figure 4. Typical Section Page.

specified assessment rate to cover such expenses will tend to effectuate the declared policy of the Act.

It is further found that good cause exists for not postponing the effective date of this action until 30 days after publication in the **Federal Register** [5 U.S.C. 553] because the Committee needs to have sufficient funds to pay its expenses which are incurred on a continuous basis. The 1994–95 fiscal year for the program began July 1, 1994. The marketing order requires that the rate of assessment apply to all assessable papayas handled during the fiscal year. In addition, handlers are aware of this action which was recommended by the Committee at a public meeting and published in the **Federal Register** as an interim final rule. No comments were received concerning the interim final rule that is adopted in this action as a final rule without change.

List of Subjects in 7 CFR Part 928

Marketing agreements, Papayas, Reporting and recordkeeping requirements.

For the reasons set forth in the preamble, 7 CFR part 928 is amended as follows:

PART 928—PAPAYAS GROWN IN HAWAII

Accordingly, the interim final rule amending 7 CFR part 928 which was published at 59 FR 33898 on July 1, 1994, is adopted as a final rule without change.

Dated: August 25, 1994.

Eric M. Forman,

Acting Deputy Director, Fruit and Vegetable Division.

[FR Doc. 94–21636 Filed 8–31–94; 8:45 am]

BILLING CODE 3410–02–P

7 CFR Part 947

[Docket No. FV94–947–2FIR]

Oregon-California Potatoes; Expenses and Assessment Rate

AGENCY: Agricultural Marketing Service, USDA.

ACTION: Final rule.

SUMMARY: The Department of Agriculture (Department) is adopting as a final rule, without change, the provisions of an interim final rule that authorized expenses and established an assessment rate that will generate funds to pay those expenses. Authorization of this budget enables the Oregon-California Potato Committee (Committee) to incur expenses that are reasonable and necessary to administer the program. Funds to administer this program are derived from assessments on handlers.

EFFECTIVE DATES: July 1, 1994, through June 30, 1995.

FOR FURTHER INFORMATION CONTACT: Martha Sue Clark, Marketing Order Administration Branch, Fruit and Vegetable Division, AMS, USDA, P.O. Box 96456, room 2523–S, Washington, DC 20090–6456, telephone 202–720–9918, or Teresa L. Hutchinson, Northwest Marketing Field Office, Fruit and Vegetable Division, AMS, USDA, Green-Wyatt Federal Building, room 369, 1220 Southwest Third Avenue, Portland, OR 97204, telephone 503–326–2724.

SUPPLEMENTARY INFORMATION: This rule is issued under Marketing Agreement No. 114 and Order No. 947, both as amended (7 CFR part 947), regulating the handling of Irish potatoes grown in Oregon-California. The marketing agreement and order are effective under the Agricultural Marketing Agreement Act of 1937, as amended (7 U.S.C. 601–674), hereinafter referred to as the Act.

The Department is issuing this rule in conformance with Executive Order 12866.

This rule has been reviewed under Executive Order 12778, Civil Justice Reform. Under the marketing order now in effect Oregon-California potato handlers are subject to assessments. Funds to administer the Oregon-California potato order are derived from such assessments. It is intended that the assessment rate as issued herein will be applicable to all assessable potatoes during the 1994–95 fiscal period, which began July 1, 1994, and ends June 30, 1995. This final rule will not preempt any State or local laws, regulations, or policies, unless they present an irreconcilable conflict with this rule.

The Act provides that administrative proceedings must be exhausted before parties may file suit in court. Under section 8c(15)(A) of the Act, any handler subject to an order may file with the Secretary a petition stating that the order, any provision of the order, or any obligation imposed in connection with the order is not in accordance with law and requesting a modification of the order or to be exempted therefrom. Such handler is afforded the opportunity for a hearing on the petition. After the hearing the Secretary would rule on the petition. The Act provides that the district court of the United States in any district in which the handler is an inhabitant, or has his or her principal place of business, has jurisdiction in equity to review the Secretary's ruling on the petition, provided a bill in equity is filed not later than 20 days after the date of the entry of the ruling.

Pursuant to the requirements set forth in the Regulatory Flexibility Act (RFA), the Administrator of the Agricultural Marketing Service (AMS) has considered the economic impact of this rule on small entities.

The purpose of the RFA is to fit regulatory actions to the scale of business subject to such actions in order that small businesses will not be unduly or disproportionately burdened. Marketing orders issued pursuant to the Act, and the rules issued thereunder, are unique in that they are brought about through group action of essentially small entities acting on their own behalf. Thus, both statutes have small entity orientation and compatibility.

There are approximately 550 producers of Oregon-California potatoes under this marketing order, and approximately 40 handlers. Small agricultural producers have been defined by the Small Business Administration (13 CFR 121.601) as those having annual receipts of less than $500,000, and small agricultural service firms are defined as those whose annual receipts are less than $5,000,000. The majority of Oregon-California potato producers and handlers may be classified as small entities.

The budget of expenses for the 1994–95 fiscal period was prepared by the Oregon-California Potato Committee, the agency responsible for local administration of the marketing order, and submitted to the Department for approval. The members of the Committee are producers and handlers of Oregon-California potatoes. They are familiar with the Committee's needs and with the costs of goods and services in their local area and are thus in a position to formulate an appropriate budget. The budget was formulated and discussed in a public meeting. Thus, all directly affected persons have had an opportunity to participate and provide input.

The assessment rate recommended by the Committee was derived by dividing anticipated expenses by expected shipments of Oregon-California potatoes. Because that rate will be applied to actual shipments, it must be established at a rate that will provide sufficient income to pay the Committee's expenses.

The Committee unanimously recommended a budget of $45,100, $1,500 more than last season. Increases in expenditures, which include $150 for the Committee's annual report, $50 for the Committee's audit, $1,000 for inspection fees, $500 for investigation

Figure 5. Typical Detail Page.

Reader Aids

Federal Register
Vol. 59, No. 95
Wednesday, May 18, 1994

INFORMATION AND ASSISTANCE

Federal Register

Index, finding aids & general information	202–523–5227
Public inspection announcement line	523–5215
Corrections to published documents	523–5237
Document drafting information	523–3187
Machine readable documents	523–3447

Code of Federal Regulations

Index, finding aids & general information	523–5227
Printing schedules	523–3419

Laws

Public Laws Update Service (numbers, dates, etc.)	523–6641
Additional information	523–5230

Presidential Documents

Executive orders and proclamations	523–5230
Public Papers of the Presidents	523–5230
Weekly Compilation of Presidential Documents	523–5230

The United States Government Manual

General information	523–5230

Other Services

Data base and machine readable specifications	523–3447
Guide to Record Retention Requirements	523–3187
Legal staff	523–4534
Privacy Act Compilation	523–3187
Public Laws Update Service (PLUS)	523–6641
TDD for the hearing impaired	523–5229

ELECTRONIC BULLETIN BOARD

Free **Electronic Bulletin Board** service for Public Law numbers, Federal Register finding aids, and list of documents on public inspection. 202–275–0920

FAX-ON-DEMAND

The daily Federal Register Table of Contents and the list of documents on public inspection are available on the National Archives fax-on-demand system. There is no charge for the service except for long distance telephone charges the user may incur. 301–713–6905

FEDERAL REGISTER PAGES AND DATES, MAY

22491–22722.....2
22723–22950.....3
22951–23118.....4
23119–23610.....5
23611–23788.....6
23789–24028.....9
24029–24340.....10
24341–24630.....11
24631–24884.....12
24885–25286.....13
25287–25554.....16
25555–25774.....17
25775–26096.....18

CFR PARTS AFFECTED DURING MAY

At the end of each month, the Office of the Federal Register publishes separately a List of CFR Sections Affected (LSA), which lists parts and sections affected by documents published since the revision date of each title.

3 CFR

Proclamations:
6599 (Revoked by Proc. 6685).....24337
6679.....22955
6680.....22957
6681.....22959
6682.....23611
6683.....24027
6684.....24335
6685.....24337
6686.....24627
6687.....24629
6688.....24883
6689.....25783
Administrative Orders:
Memorandums:
April 29, 1994.....22951
April 29, 1994.....22953
Presidential Determinations:
94–23 of May 3, 1994.....24625
Executive Orders:
8685 (Revoked in part by PLO 7045).....24947
12582 (Revoked by EO 12913).....23115
12779 (See EO 12914).....24339
12866 (See OMB report of May 1).....24276
12878 (Amended by EO 12912).....22949
12912.....22949
12913.....23115
12914.....24339
12915.....25775
12916.....25779
1919½ (Revoked in part by PLO 7046).....24648
3406 (Revoked in part by PLO 7048).....24649

5 CFR

531.....24029
890.....24030
Proposed Rules:
Ch. XIV.....22537
185.....24661
213.....25966
630.....24560
890.....24062

7 CFR

58.....24318
80.....25785
91.....24318
93.....24318
94.....24318
95.....24318
98.....24318
210.....23613
220.....23613
271.....22723
272.....22723
273.....22723
301.....22491, 25287, 25786, 25789
723.....22723
944.....25791
956.....24631
982.....24632
993.....25792
998.....24633
1033.....24030
1036.....24030
1049.....24030
1240.....22492, 24217
1413.....22494, 22495, 25794
1421.....25794, 25795
1427.....22494, 22495
1434.....23789
1464.....22723
1540.....25796
1941.....22961, 25794, 25797
1943.....22961, 25797
1945.....22961, 25797
1948.....24635
1951.....25797
1980.....23614
Proposed Rules:
Ch. VII.....22938
201.....25706
300.....22538, 24968
319.....22538, 24968
58.....24318
993.....25841
999.....25841
1468.....22546
1530.....23017
1823.....23018
1910.....23018
1927.....24362
1941.....23018
1942.....23018
1943.....23018
1944.....22018
1945.....23018
1948.....23018
1951.....22548, 23018
1980.....23018, 23173
2812.....24973
4285.....23804

8 CFR

210a.....24031
299.....25555
499.....25555
Proposed Rules:
3.....24976, 24977
245a.....24978

9 CFR

78.....22496

Figure 6. Typical Appendix Page.

3.0 IMAGE SCANNING

Before scanning each *Federal Register* book, its pages were cut from its binding. This resulted in paper pages that were approximately 20 CM (8") by 28 CM (11") in size. The NIST *Federal Register* collection is missing one issue for the 1994 calendar year: March 10. The collection, containing 255 books, was scanned by a contractor at its Rockville, MD facility. Image scanning was performed using a Kodak 923[1] scanner at a resolution of 15.75 pixels per millimeter (400 pixels per inch (PPI)) to output a compressed bitonal tagged image formatted file(TIFF™[8]). Image files were written to CD Recordable cartridges (CDR) using an image file naming convention, "/Nmm/nn/nn/nn/nn.tif" where:

- mm denotes the CDR volume where mm = 01..23
- nn denotes a path or filename where nn = 0..99

This naming convention allows for storage of 100 image files within each sub-directory entity. For example, the image file "N18/00/00/00/00.tif" contains page number 44606. Page number 44607 is contained on the next sequentially numbered file. This numbering convention continues with image file "N18/00/00/01/00.tif" containing page number 44707.

Approximately, 67,000 *Federal Register* pages were scanned and the resultant TIFF™ images were written to 22 CDR cartridges (CDR cartridge N20 was skipped). Figure 7 illustrates the GPO printing volumes in terms of pages per month.

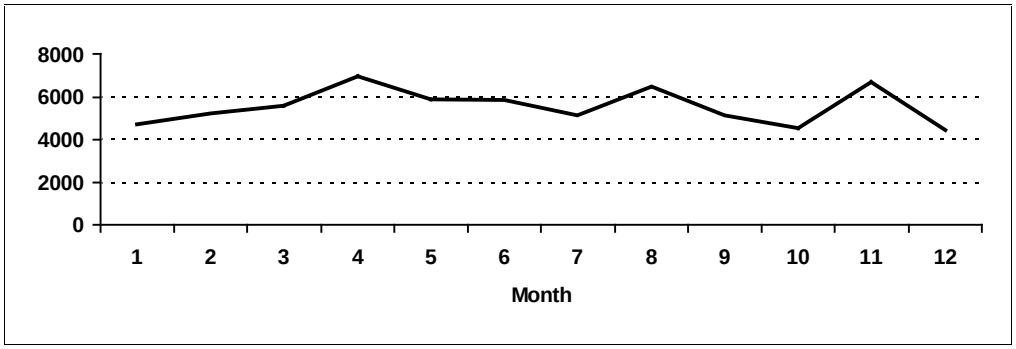


Figure 7. Monthly Printing Volumes for 1994 of the Federal Register.

4.0 IMAGE VALIDATION PROCESSING

Figure 8 illustrates the general process flow that was used to map and validate the CDR images from the above serial structure, "/Nmm/nn/nn/nn/nn.tif", into a mapping structure which uses a month, day, and page number convention. This process translated the filenames, assured image quality, and verified the correct page was stored. It consisted of Sections 4.1 Image Name Mapping, 4.2 Image Quality Checks, 4.3 Arabic Page Numbered Image Verification, 4.4 Roman Page Numbered Image Verification, and 4.5 Image Sequence Check Verification.

---

[1] Specific hardware and software products identified in this paper were used in order to adequately support the development of the technology described in this document. In no case does such identification imply recommendation or endorsement by NIST, nor does it imply that the equipment identified is necessarily the best available for the purpose.

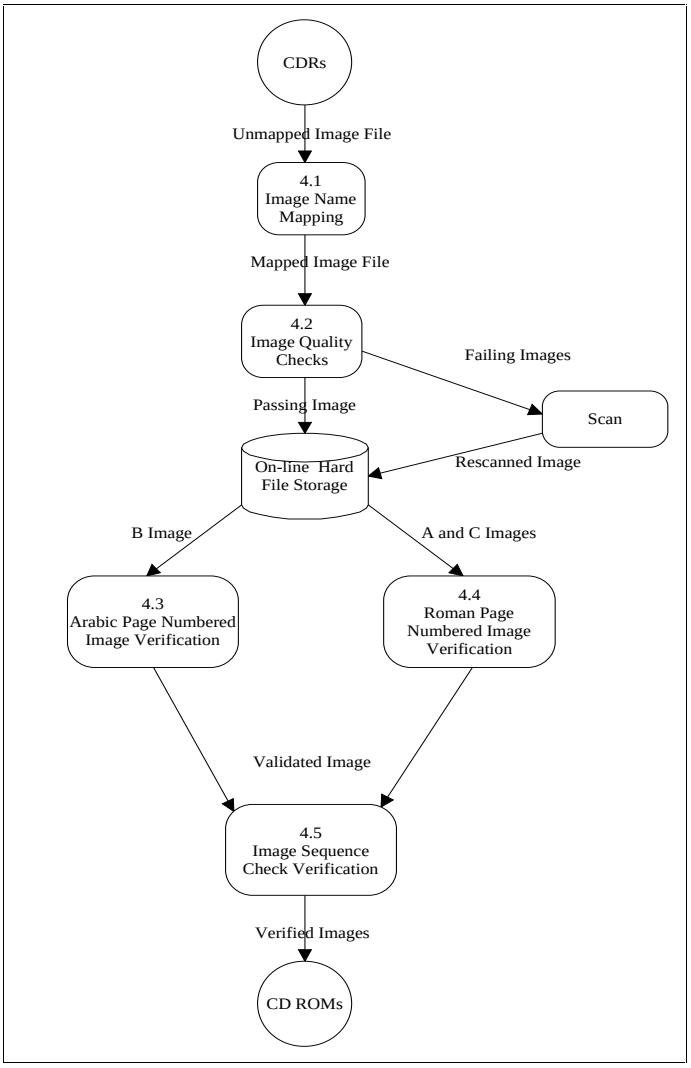


Figure 8. Top Level Federal Register Validation Process Flow.

4.1 Image Name Mapping

As each image file was read from CDR, the path and filename was mapped into a name consisting of a two-digit month subdirectory name, a two-digit day subdirectory name, and an eight character filename suffixed with a three-character filename extension ("/mm/dd/filename.ext" where mm = 1..12 and dd = 1..31). Daily, the GPO generates Microcomp files, for each *Federal Register* book, that contain information required to perform this mapping. We created a mapping index and verified that it was correct by manually checking it against each *Federal Register* book.

Each eight-character filename conforms to the convention of "t00nnnnn" where "t" denotes the type of *Federal Register* page and "00nnnnn" represents a page number. "nnnnn" is zero filled and padded to the left. The file naming conventions were:

- Prefix page image filenames were assigned "t = a"; "00nnnnn" was reset to zero for each book. The hard cover page was named "a0000000" and the soft cover page was named "a0000001". Each subsequent prefix page was assigned a page number that was sequentially incremented.

- Body page image filenames were assigned “t = b”; “00nnnnn” was sequentially incremented for the entire year and is the actual *Federal Register* page number.
- Appendix page image filenames were assigned “t = c”; “00nnnnn” was reset for each book. The initial appendix page was named “c0000001”. Each subsequent appendix page was assigned a page number that was sequentially incremented.

4.2 Image Quality Checks

Next, each image file was converted from TIFF™ to NIST IHead [9] format. During this conversion, truncated and/or corrupted bitmaps generated by the scanning process were detectable. If a bad image bitmap was detected, the image name was flagged and its original *Federal Register* page was rescanned. If not, a three-character “pct” extension was assigned and the file was stored upon on-line magnetic file storage.

In order to ensure an image file was ready for further verification processing, a series of image quality checks were performed to ensure that each file conformed to the following characteristics:

- Resolution = 15.75 pixels per millimeter (400 PPI)
- Compressed file size of CCITT Group 4 [10] image ≥ 30 K Bytes (KB)
- Width < 4000 pixels
- 4200 pixels < Height ≤ 4900 pixels

An image file that did not conform to these characteristics was flagged and the original *Federal Register* page was rescanned.

We decided not to store any images of blank *Federal Register* pages on the output media. All compressed image files of less than 30 KB were visually inspected, verified to be blank within the *Federal Registe*r paper book, and deleted from on-line magnetic storage. However, the image file sizes of several blank pages were greater than the 30 KB criteria (due to excessive speckling) and were eliminated during subsequent validation procedures.

Figure 9 illustrates the rescanning attributable to image quality problems that was detected by the above quality thresholds. It does not include 1362 image rescans, occurring predominately in month “7”, due to an incorrect scanning resolution of 11.81 pixels per millimeter (300 PPI). In total, approximately 1790 pages were rescanned which represented 2.7% of the total number of pages. All rescanning was performed in our laboratory using a Fujitsu 3096G[1] scanner.

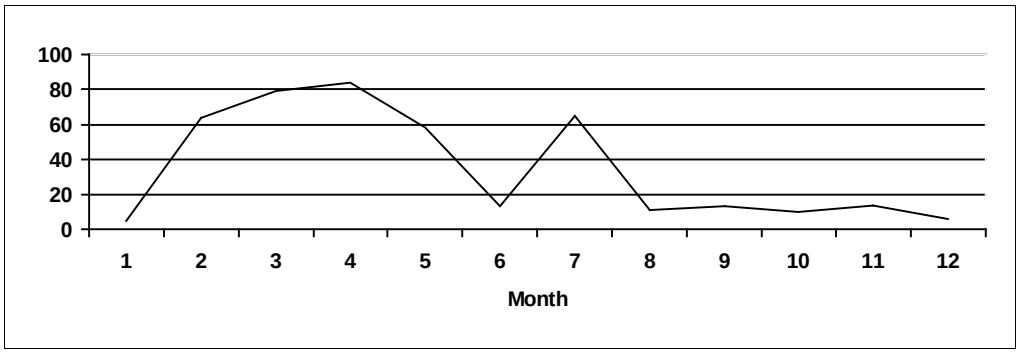


Figure 9. Rescans Due to Image Quality.

4.3 Arabic Page Numbered Image Verification

Figure 10 illustrates the process steps that validated each body page image of the *Federal Register*, "b00nnnnn". It consisted of Sections 4.3.1 Locate Page Number and Create Subimage, 4.3.2 Optical Character Recognition of Page Number, 4.3.3 Full-Page Image Adjudication, 4.3.4 Compare OCR Results to Image Filename, and 4.3.5 Subimage Adjudication using Page Number.

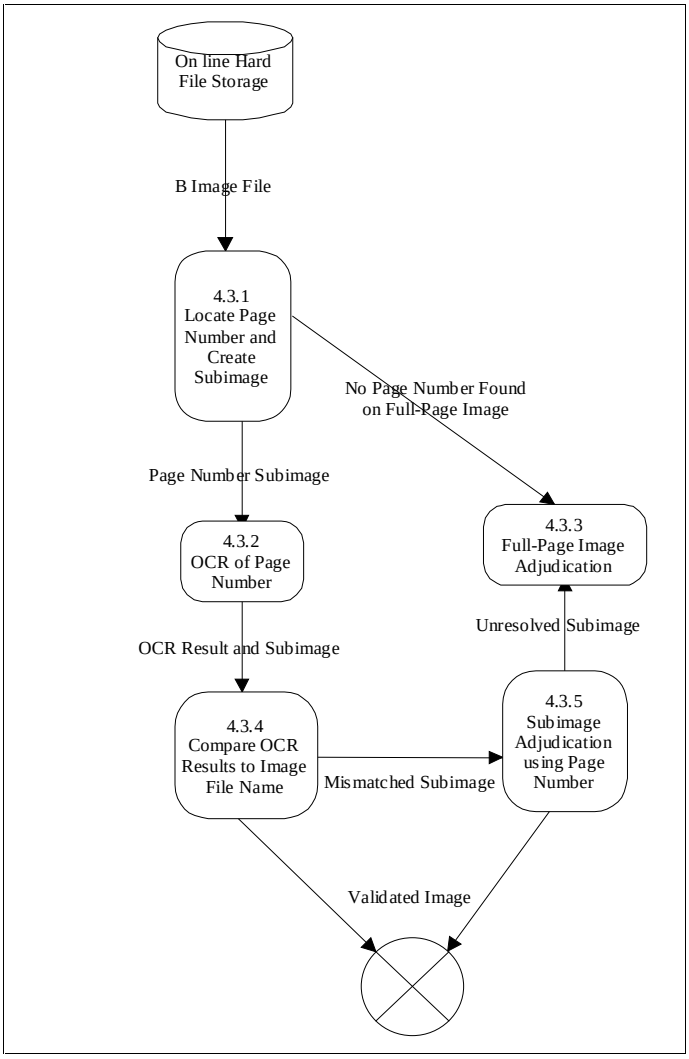


Figure 10. Arabic Numbered Type of Page Processing.

4.3.1 Locate Page Number and Create Subimage

Figure 5 illustrates the page heading of a *Federal Register* body type of page which contains a page number field. Depending upon the page face, the page number will be printed on either the left edge of the heading when it is even numbered or right edge of the heading when it is odd numbered. Each image was decompressed and a subimage of the top 5 CM (800 pixels) was created from the raster image. The subimage was spatially reduced by a factor of 3 to increase efficiency, a skew angle was computed[11]. If the subimage was skewed by more than 0.2 degrees, it was rotated and the skew was removed. Then, the horizontal header lines, below the text banner, were located and the subimage was truncated to exclude them.

Isolation of the page number field was attempted by examining the text banner line and locating the page number field on either the right or left side. If this isolation was unsuccessful, the failing full-page image was sent to Section 4.3.3 for adjudication. If successful, a subimage file containing the page number was created and was input to the recognition processing.

Figure 11 provides a graphical analysis of page number isolation failures that include measurements of blank, section, and skewed pages. Figure 12 illustrates an example of a subimage (Width = 192 pixels and Height = 53 pixels) that has been enlarged by 150%.

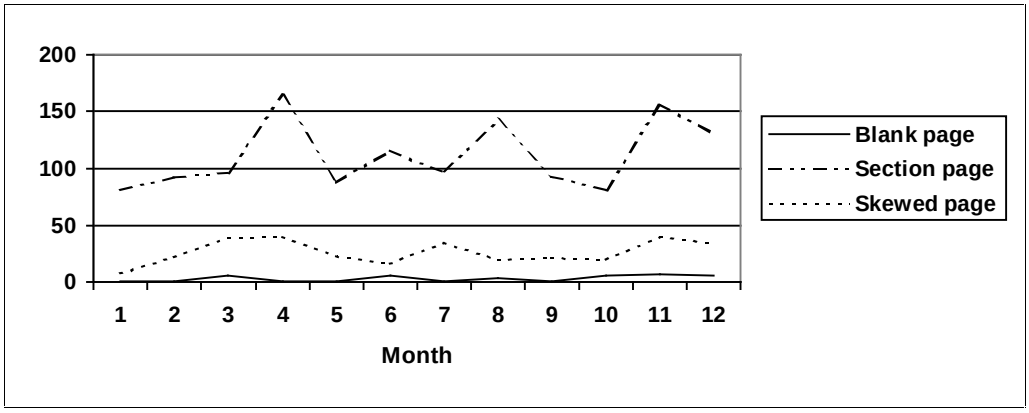


Figure 11. Page Number Isolation Failures.

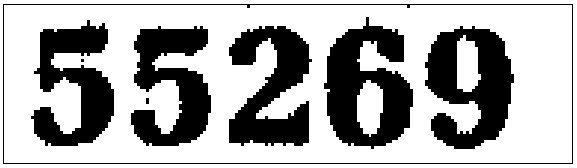


Figure 12. Typical Subimage.

4.3.2 Optical Character Recognition (OCR) of Page Number

Since over 95% of the *Federal Register* consists of body type of pages, NIST decided to use OCR rather than human inspection to validate that each candidate image contained the correct *Federal Register* page, and that it was in its correct position within the month and day directory structure.

Initially, we examined and tested page number recognition with three commercially available OCR products that executed in Microsoft Windows/NT and UNIX environments. Accuracy from each of the three products was determined to be unsatisfactory to NIST because the amount of touching characters contained within the *Federal Register* yielded a low omni-font recognition accuracy.

As a result of this evaluation, NIST decided to use it own OCR capabilities to recognize the digits contained within the page number subimage. This involved the following activities:

4.3.2.1 Segmentation

Our previous releases of the NIST public domain Form-Based Handprint Recognition System [9] (HSFSYS) contained segmentation software that isolated handprinted characters. This software was modified and adapted to segment 9-point Vermilion page number machine printed digits. Figure 13 and

Figure 14 illustrate typical *Federal Register* page number fields that contain digits which form touching characters. Our handprint segmentation software was modified to properly segment machine printed characters similar to these examples.

Figure 13. Touching Characters caused by too much ink.

Figure 14. Touching Characters caused by ink bleeding through page.

4.3.2.2 Classification and Training

HSFSYS Release 1.0 contains software that uses a Probabilistic Neural Network (PNN) [12] to classify a segmented character. Release 2.0 contains software that added a Multi–Layered Perceptron (MLP) [13] to classify a segmented character. For this application, a PNN classifier was chosen to classify segmented characters on the basis of taking less effort to train than an equivalent MLP classifier. The PNN classifier was trained to recognize 10 digits, 0-9. Due to time constraints, the classifier was trained using only 100 discrete samples per digit class. We know that a larger training set would improve recognition accuracy.

4.3.2.3 OCR System Accuracy

For 64,384 subimages, our system accuracy for OCR of the page number field achieved an 88.1% overall correct recognition. System accuracy includes page number isolation errors as well as typical OCR segmentation and classification errors. Figure 15 presents a graphical view of each month's accuracy. The variance in bleed through, smudged, and lightly printed ink conditions provided an extremely difficult recognition challenge.

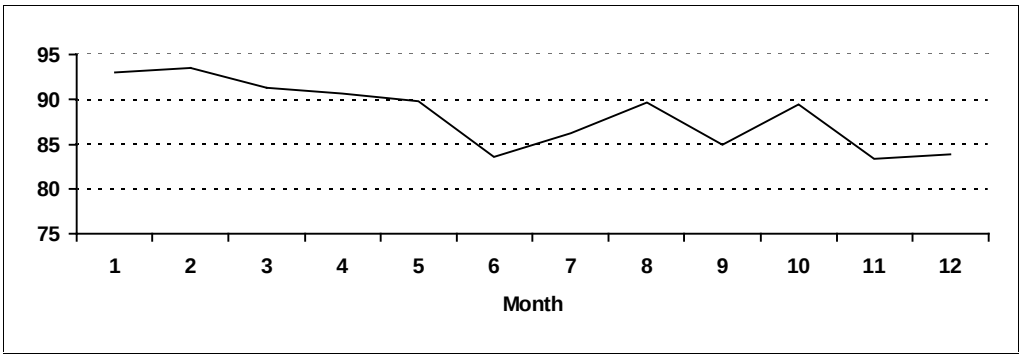


Figure 15. OCR System Accuracy on Page Number Fields.

Over 40% of the OCR errors are attributable to improper character segmentation of the page number subimages; this is an unusually high percentage of failure caused by touching and incomplete/broken characters. The remaining errors are attributable to improper classifications by our PNN classifier.

Our OCR accuracy varied greatly with the quality of the printed material. The GPO printed the *Federal Register* on recycled newspaper quality of paper, highly absorbent, using a printing plate that often contained either too much ink or too little ink. As illustrated in Figure 16, too much ink and/or bleed through conditions resulted in failures by our segmentation software to correctly segment three or more touching characters.

As illustrated in Figure 17, too little ink resulted in missing digit(s) and/or broken character segments. Although our software correctly segmented and classified a subimage containing missing character(s), it scored the result as an OCR error in the above accuracy graph because the OCR result did not exactly match the filename string (Section 4.3.3.3). As illustrated in Figure 18, the ink is lightly printed. At times, our segmentation software cut the non-contiguous sections of characters into multiple characters. Both of these conditions were major contributors to the number of segmentation errors.

Figure 16. Segmentation Error caused by Touching Digits.

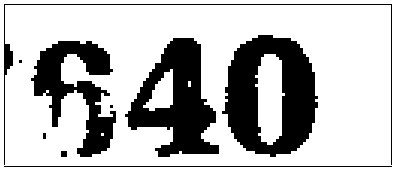


Figure 17. OCR Error caused by Missing Printed Digit(s).

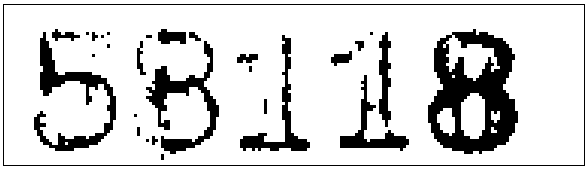


Figure 18. Segmentation Error caused by Light Ink Printing.

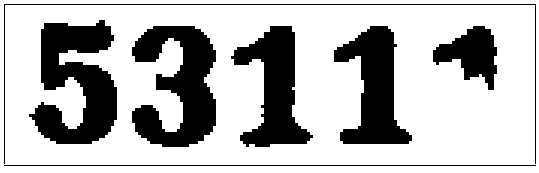


Figure 19. Classification Error caused by Incomplete Training Set.

Figure 20 illustrates a confusion matrix that was generated by examining all mismatched recognition result strings, 4269 occurrences, which contained the same number of digits as its associated truth (filename) string. Although it does not completely eliminate segmentation errors from the analysis, it does provide an interesting view of false classifications by our PNN.

Digit	**Recognized As**									
	0	1	2	3	4	5	6	7	8	9
0	0	24	7	13	78	7	60	5	510	139
1	2	0	14	34	27	2	1	49	5	8
2	1	78	0	82	4	7	13	32	22	14
3	4	200	81	0	19	34	16	123	105	28
4	2	103	1	19	0	6	4	13	13	10
5	2	66	4	98	58	0	146	40	140	12
6	34	36	1	11	487	27	0	5	1280	16
7	2	54	34	1	4	8	13	0	11	11
8	4	40	8	41	56	16	80	9	0	12
9	8	41	3	55	33	10	107	17	160	0

Figure 20. Confusion Matrix.

A number of observations can be drawn from the above confusion matrix. Fifty five percent of the total number of errors in the matrix are represented by the following six confusion pairs:

- “6” classified as an “8” (30%)
- “0” classified as an “8” (12%)
- “5” classified as a “6” (3.4%)
- “5” classified as an “8” (3.3%)
- “9” classified as an “8” (3.8%)
- “3” classified as an “8” (2.5%)

Upon visual inspection, it was determined that ink bleed through was the primary source of error. In addition, sixteen percent of the total number of errors in the matrix are represented by the following confusion pairs:

- “3” classified as a “1” (4.6%)
- “6” classified as a “4” (11.4%)

These errors are primarily attributed to the limited size of character samples used. Training the PNN with 100 character samples per class is not sufficient. Recognition accuracy can be improved by training on a

larger number of diversified samples. Figure 19 provides an illustration of this condition; our character training sets did not include enough distinctive partial characters to classify incomplete characters.

4.3.3 Full-Page Image Adjudication

Whenever our page number isolation software failed or the visual subimage adjudication was not successful, full-page image adjudication processing was performed by an operator using a 2.6 pixels per millimeter (66 PPI) reduced full-page image. As illustrated in Figure 21, it allowed an operator to adjudicate blank and non-blank page images.

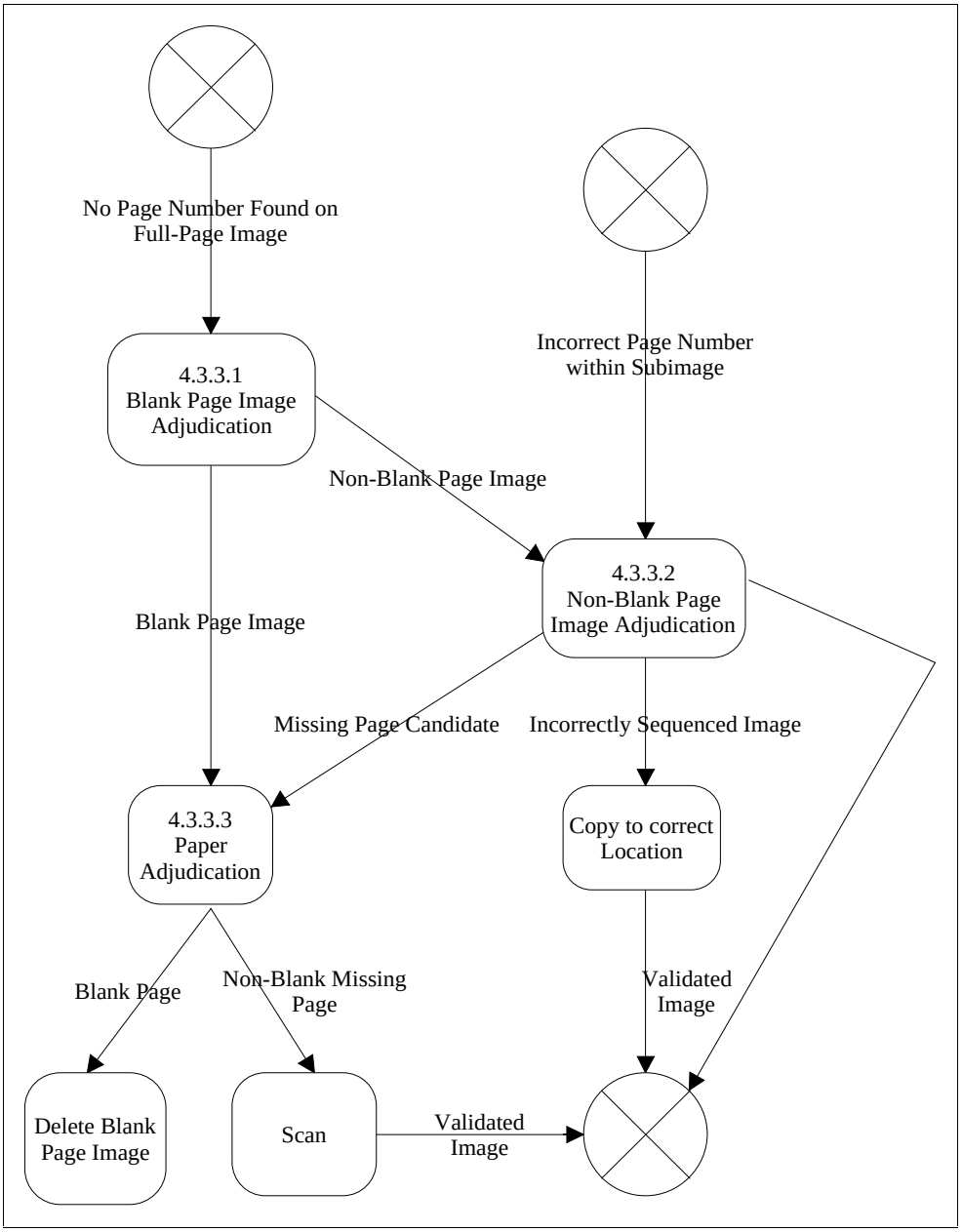


Figure 21. Full-Page Image Adjudication.

4.3.3.1 Blank Page Image Adjudication

Whenever our page number isolation software, described in 4.3.1, failed to locate and isolate a page number, full-page images were routed to this process step. An operator viewed each candidate and classified the image as either a blank or non-blank page.

- A blank page: Due to speckling, several blank page images were encountered because their file sizes exceeded the 30KB image quality threshold.
- A non-blank page: A section page within the body of the *Federal Register* does not contain a page number header and failed the isolation processing. Extremely skewed pages also failed this processing. Upon visual inspection, those images that contained more than ten degrees of skew were rescanned. For each page, an operator confirmed that it was in correct sequence by checking the *Federal Register* book. If not, the incorrectly sequenced image was copied to its correct location and a missing page candidate was flagged.

Blank page and missing page candidates required adjudication using the actual printed *Federal Register* page, Section 4.3.3.3.

4.3.3.2 Non-Blank Page Image Adjudication

Failures were routed to this stage by an operator (1) identifying a non-blank page for which no page number was isolated or (2) failing to adjudicate a subimage from the information content. At this stage, an operator categorized whether or not a page was:

- Scanned in the wrong sequence.
- Printed with missing or illegible digit(s)

An incorrectly sequenced image was copied to its correct location and a missing page candidate condition was flagged. The missing page candidate was adjudicated using the same procedures for adjudication of non-blank pages described in Section 4.3.3.3.

At times the page number isolation software isolated and created a subimage that an operator could not validate from the information contained within the subimage. Whenever a page number contained within a subimage was not readable by an operator, the image was verified to be correct by comparing the full-page image content with the printed paper page information content. Figure 22 illustrates this case; the low order digit is absent from the printed *Federal Registe*r page. Only examination of the printed page content revealed whether or not the image file was the actual scanned paper page.

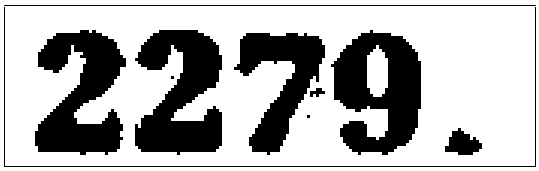


Figure 22. Image Snippet contained Unprinted Low Order Page Number Digit.

4.3.3.3 Paper Adjudication

Blank pages from Section 4.3.3.1 and missing page candidates from Section 4.3.3.2 required adjudication by an operator using the paper pages of the corresponding *Federal Register book*. If an actual paper page

was verified to be blank, the full-page image was deleted from within the "mm/dd" directory hierarchy. If non-blank, the original *Federal Register* page was missing and was scanned.

4.3.4 Compare OCR Results to Image Filename

Characters in each page number subimage were recognized and classified by the NIST OCR software. It output two files containing recognition results: hypothesis strings and confidence values. For each character that was segmented and classified, the hypothesis file contained an ASCII character and the confidence file contained a value that represented the confidence of correct character classification. The process we used to validate images did not use the recognition confidence values. Instead, we chose to rely upon an exact match of the ASCII OCR results with the numeric portion of its filename.

For each page number subimage, its filename string (truncated to exclude leading zeros) was compared with the OCR results string contained within its associated hypothesis file. If an exact string match was made, the image was assumed to contain the correct *Federal Register* page and that it was correctly stored within the "mm/dd" directory hierarchy. If the strings did not match, the file was classified as a mismatched subimage that required adjudication by a key entry operator.

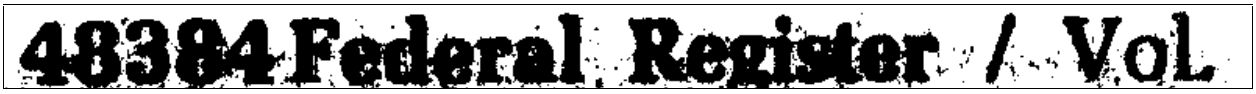


Figure 23. Page Number Isolation Error.

4.3.5 Subimage Adjudication using Page Number

At times the page number isolation failed due to excessive skew, too much noise, or text was printed too close to the page number. An example of the latter is illustrated in Figure 23; the word "Federal" abuts the page number field. The OCR results contained correct character classifications for the page number digits and erroneous classifications for the non-numeric characters. We could have implemented data validation software that could have detected this condition and determined that there was a matching condition; however, we chose to rely on a human being to adjudicate this type of failure.

Page number subimages were amassed for high-speed human adjudication. Using a high resolution visual display terminal, an operator was presented with (1) a window that displayed the ASCII image filename in its title area and a subimage in the display area and (2) another window that allowed an operator to respond and verify whether or not the page number was either correct or incorrect.

After viewing the page number subimage and visually reading and comparing it with the filename string displayed in the title area of its window, an operator responded with single keystroke of either "y" or "n". If an operator accepted the page number, it was assumed that the page was scanned correctly and the image file was correctly positioned within the "mm/dd" directory hierarchy. If not, the image file required further adjudication.

4.4 Roman Page Numbered Image Verification

Figure 24 illustrates the process steps which validated the prefix and appendix pages, "a00000nn" and "c00000nn", of the *Federal Register*. It consisted of Sections 4.4.1 Separate Cover Pages from Content Pages and 4.4.2 Cover Page Image Adjudication.

4.4.1 Separate Cover Pages from Content Pages

Since cover page image files (“a0000000” and “a0000001”) do not contain page numbers, these files were separated and sent to an operator for adjudication.

All non-cover page (“a0000002” through “A00000nn” and “c00000nn”) image files were verified by using process steps similar to the ones described in Sections 4.3.1, and 4.3.5.

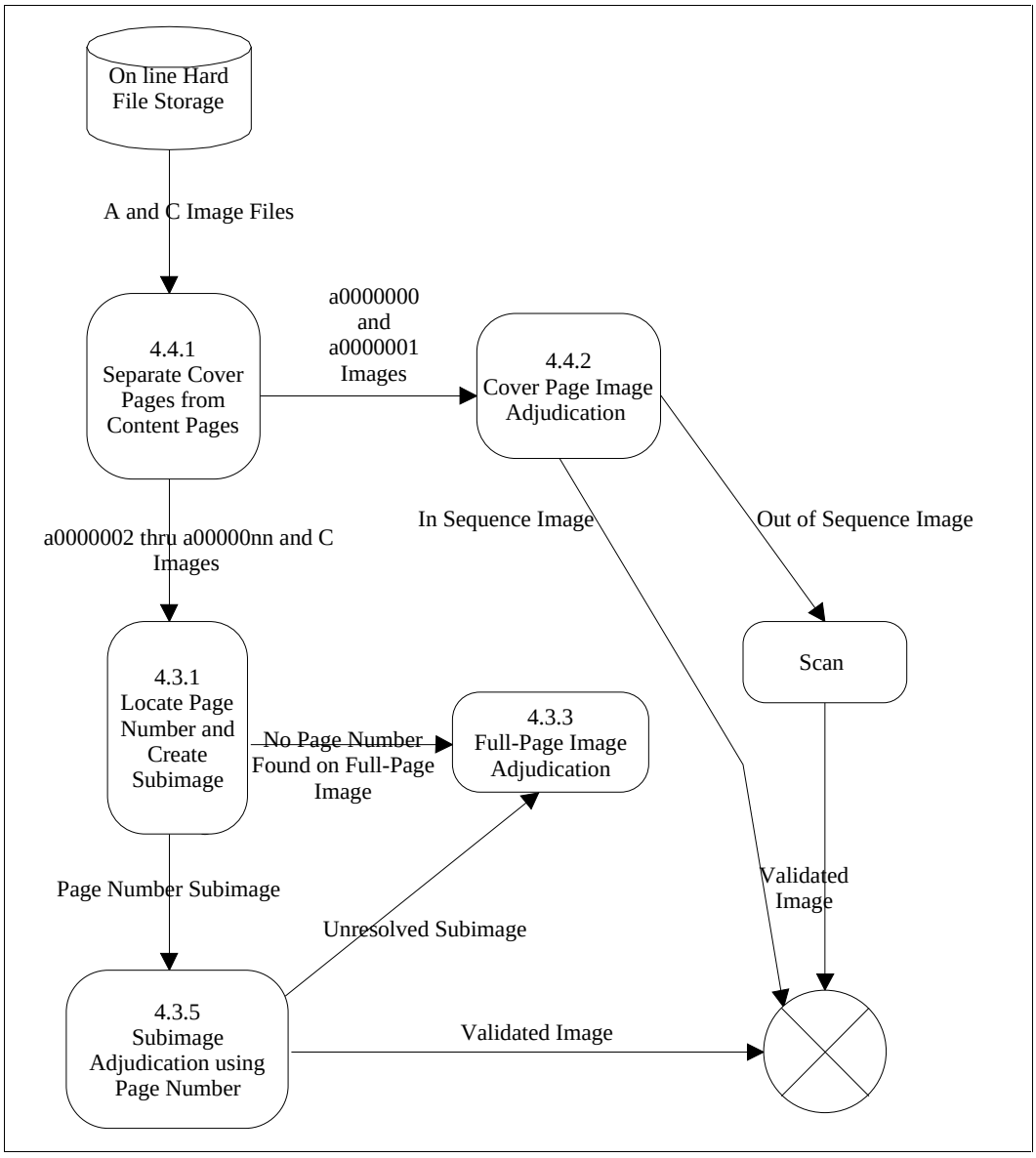


Figure 24. Roman Numbered Type of Page Processing.

4.4.2 Cover Page Image Adjudication

Each hard cover page ("a0000000") and soft cover page ("a0000001") was viewed by an operator at a high resolution visual display terminal, displaying a 2.6 pixels per millimeter (66 PPI) reduced full-page image. Since these pages are distinctive (Figure 1 and Figure 2), an operator quickly confirmed whether or not these pages were correctly stored within the "mm/dd" directory hierarchy. If not, an out of sequence condition was found and the original *Federal Register* page was rescanned.

4.5 Image Sequence Check Verification

The final step within this verification process entailed performing an image sequence check that detected and examined page numbering gaps within the sequence of image files within the "mm/dd" directory hierarchy. Each gap was either explained or corrected. Numerical gaps were caused by either failing to scan *Federal Register* page(s), blank pages detected and deleted by the validation processing, or legitimate page number increments made by the GPO. An operator adjudicated these conditions by reviewing processing logs and/or verifying the gap condition within the *Federal Register*.

5.0 SUMMARY

We realize that the cost of scanning and verifying any voluminous document collection, such as the *Federal Register,* is a tedious and expensive process. It cost $8,000 to scan the approximately 67,000 pages of the *Federal Register*. In order to reduce the verification cost, we decided to minimize human resource expenditure by exploiting OCR and high-speed visual adjudication by an operator. Usage of OCR allowed us to automatically validate and exclude over 83% of the images from being adjudicated by a human. Of the remaining 17% of images, over 90% of these images were validated by high-speed operator adjudication; this minimized expensive paper handling.

Even though the validation was semi-automated, it was conducted by highly skilled professionals at NIST and required a one-person month of labor costing approximately $35,000. We believe that certain subjective judgements are best made by technically oriented image processing professionals. If a lesser labor category were substituted, the cost of validation could be reduced at the expense of quality.

6.0 REFERENCES

[1] C.L.Wilson, J. Geist, M.D. Garris, and R. Chellappa, "Design, Integration, and Evaluation of Form-Based Handprint and OCR Systems," NIST Internal Report 5932, December 1996.

[2] D. Harman, "Towards Interactive Query Expansion," Proceedings of the 11TH International Conference on Research & Development in Information Retrieval (SIGIR – 88), pp. 7-15, June, 1988.

[3] D. Harman, "How Effective is Suffixing?," Journal of the American Society for Information Science(JASS), Vol. 42, #1, pp. 7-15, 1991.

[4] D. Harman, "Analysis of Data from Second Text Retrieval Conference (TREC – 2)," Proceedings of RIAO-94, pp. 699-709, October, 1994.

[5] R.A.Wilkinson, J. Geist, S. Janet, P.J. Grother, C.J.C. Burges, R. Creecy, B. Hammond, J.J. Hull, N.J. Larsen, T.P. Vogel, and C.L. Wilson, " The First Census Optical Character Recognition Conference," NIST Internal Report 4912, August 1992.

[6] J. Geist, R.A. Wilkinson, S. Janet, P.J. Grother, B. Hammond, N.W. Larsen, R.M. Klear, M.J. Matsko, C.J.C. Burges, R. Creecy, J.J. Hull, T.P. Vogel, and C.L. Wilson, "The Second Census Optical Character Recognition Conference," NIST Internal Report 5452, May 1994.

[7] D. Harman Ed., "The Fourth Text Retrieval Conference (TREC-4)," NIST Special Publication 500-236, November 1995.

[8] "TIFF™, Revision 6.0," June 3, 1992, Aldus Corporation, Seattle, WA 98104-2871.

[9] "NIST Form-Based Handprint Recognition System (Release 2.0)," NISTIR 5959, January 1997.

[10] "Facsimile Coding Schemes and Coding Control Functions for Group 4 Fascicle VII.3 – Rec. T,6," 1984, CCITT.

[11] M.D. Garris and P.J. Grother; "Generalized Form Registration Using Structure-Based Techniques," NIST Internal Report 5726 and in Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, pp. 321-334, UNLV, April 1996.

[12] D. F. Specht, "Probabilistic Neural Networks," Neural Networks, Vol. 3(1), pp. 109-119, 1990.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Parallel Distributed Processing, Volume 1: Foundations, edited by D. E. Rumelhart, J. L. McClelland, et. al., MIT Press, Cambridge, pp. 318-362, 1986.