CENDI

CENDI ANALYSIS OF SCANNING/OPTICAL CHARACTER RECOGNITION POSITION DESCRIPTIONS

Submitted by
Position Description Task Group
CENDI Information Exchange Working Group

Prepared by
Gail Hodge, CENDI Secretariat
Information International Associates, Inc.
Oak Ridge, Tennessee

January 1997

POSITION DESCRIPTION TASK GROUP PARTICIPANTS

Barbara Bauldock, Chair (DOE/OSTI)*
Charlene Luther (DOE/OSTI)*
Ken Hohenbrink (DOE/OSTI)
Doreen Owens (DOE/OSTI)
Helen Viel (DTIC)*
Lou Knecht (NLM)*
Nadgy Roey (NLM)
Larry Placanica (NTIS)*
Sue Feindt (NTIS)
Kris Vajs (NTIS)
Gail Hodge (CENDI Secretariat)

Other Government Agency Contacts

Shirley Edwards (NAL)
Carl Fleishaur (LoC)

* Indicates person established as a point of contact by the agency.

CENDI is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Commerce, Energy, Defense, Health and Human Services, Interior, and the National Aeronautics and Space Administration (NASA).

CENDI's mission is to help improve the productivity of Federal science- and technology-based programs through the development and management of effective scientific and technical information support systems. In fulfilling its mission, CENDI member agencies play an important role in helping to strengthen U.S. competitiveness and address science- and technology-based national priorities.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

At the request of the CENDI Principals (May 29, 1996), the CENDI Information Exchange Working Group organized a task group to discuss the issues related to the position descriptions for scanner and optical character recognition (OCR) operators. The purpose was to discuss changes brought about by the implementation of electronic document management system (EDMS) technologies, to review the status and appropriateness of current position descriptions, and to make appropriate recommendations.

The task group developed a survey form to collect relevant information from CENDI agencies, including position titles, grade and series information, and comments concerning changes based on the introduction of EDMS technologies. (The survey questions used to collect information are provided in Appendix I.) In addition, position descriptions were collected from the National Agricultural Library (NAL), the Library of Congress, and the Congressional Record, government agencies known to be heavily involved in text scanning. The survey information and non-CENDI position descriptions provided background material for the task group meeting held at the Department of Energy on October 1, 1996, with Barbara Bauldock, the Information Exchange Working Group Chair, moderating. Representatives from the National Library of Medicine, the Defense Technical Information Center, and the National Technical Information Service were present. In addition to Ms. Bauldock, technical representatives from the Department of Energy, Office of Scientific and Technical Information, Oak Ridge, TN, participated by videoteleconference. A list of agency representatives is provided in the front of this document.

The initial findings of the task group were presented at the CENDI Principals Meeting of October 16, 1996. Modifications to the analysis were requested by the Principals, and are included in this report.

The following observations were made:

- Agencies are generally in a state of transition with regard to the incorporation of scanner and OCR technologies. Many agencies have recently implemented or are in the process of ramping up their systems to full production.

- Agencies who have fully implemented production systems are more likely to have made changes to the position descriptions.

- Scanning and OCR technologies are used both at the initial data capture stage for incoming material and at the end of the system to convert legacy collections to electronic form or to provide output for document distribution.

- The incumbent staff most effected by the new technologies are the staff who previously performed data capture and micrographics/reprographics functions.

- The incumbents may be in several different grade series/titles. The most common involve the data entry clerks, library technicians, the micrographics specialists, and the offset printers.

- The agency that reviewed its position descriptions was able to justify changes in title, grade and series for the people using the new technologies based on the increased decision making and the increased sophistication of the decisions that need to be made.

- The changes in decisions include what resolution to use in scanning, what if any image enhancement software to use, and in some cases a decision as to whether or not the document can be scanned at all.

- Transitioning incumbents is important to successful implementation of electronic document management. Budget constraints as well as employment policies and, in some cases, hiring freezes dictated that existing staff be used.

- Position descriptions that do incorporate the new technologies identify them as tools to perform the job.

- New PDs are assigned to series depending upon on the environment in which the scanning/OCR is performed. If the environment has historically been library oriented, the PDs are in the librarian or library technician series (1410 or 1411). If the environment has been that of database creation, the PD is on the clerical series 1000. Generally, the series has not been changed but, based on the increased decision-making required with advanced scanning/OCR technologies, the agencies are implementing upgrades to the GS-5/6 range. If the Library of Congress is any indication, the addition of more complex multimedia documents may increase the range beyond the 5/6 range, even for beginning technicians.

Conclusion and Recommendations

In order to support the development of appropriate PDs, the task group presents language that can be incorporated into new or existing PDs. This language emphasizes the use of microcomputers and, for advanced positions, an understanding of scanner and OCR technologies. It also emphasizes the need to identify the potential quality of the systems output and how to adjust the equipment to correct it, or when to reject material as being unscannable.

1.0 INTRODUCTION

At the request of the CENDI Principals, the Information Exchange Working Group organized a Technology Task Group on position descriptions for personnel involved in optical scanning and optical character recognition (OCR). This project was an outgrowth of the Information Exchange Working Group's review of optical scanner and OCR technologies among the CENDI agencies. The purpose was to discuss changes brought about by the implementation of electronic document management technologies, the status and appropriateness of current position descriptions, and to make appropriate recommendations.

Prior to a formal meeting, the Senior Analyst developed a draft survey form to collect relevant information regarding current and previous position titles. Grade/series information was also collected. This survey form was reviewed and modified by the task group via e-mail. The survey forms were distributed to the representatives and the information returned to the Senior Analyst. The Senior Analyst compiled a comparison chart and brief textual analysis for presentation at the task group meeting.

The task group meeting was held on October 1, 1996 at the Department of Energy moderated by Barbara Bauldock, Information Exchange Working Group Chair. Representatives from DTIC, DOE, NLM, and NTIS were present. The DOE OSTI staff based in Oak Ridge, TN, participated by videoteleconference. The participants and their respective agencies are listed as front matter to this document. At this meeting, the task group members reviewed the data collected via the survey form, provided additional information, and made initial observations and recommendations.

The initial deliverable from this meeting was a presentation to the CENDI Principals on October 16, 1996. (A copy of the overheads is available from the CENDI Secretariat upon request.) Following the presentation, the Principals clarified some of the opinions expressed by their representatives and asked for modification to the final report. The Principals agreed with the recommendation that language which could be incorporated into new or existing PDs be developed by the Task Group.

This report reviews the information presented by each agency and the information collected from three other government agencies with digital library projects underway—the Library of Congress, the National Agriculture Library, and the Congressional Record. An attempt was made to include private sector organizations in this analysis, but no relevant position descriptions were provided. The information collected is analyzed for common problems and trends. General observations also are drawn. Suggested language is presented to use by the agencies when developing a position description for a scanner or OCR operator.

2.0 POSITION DESCRIPTIONS

2.1 CENDI Agencies

2.1.1 *Department of Energy*

There are several positions at DOE involved in the mass scanning and automatic OCR performed on incoming material. The current position titles include Printing Specialist, Offset Press Operator, Offset Photographer, and Technical Information Clerk. The position descriptions have not been changed, because the percentage of the work related to the electronic document management system performed by any one of these staff members is less than 20%. The appropriateness of the current description to the work being performed will only become a concern if the volume of work increases. When the work can no longer be performed optimally as add-on duties to the current positions, OSTI will revisit the arrangement.

2.1.2 *Defense Technical Information Center*

DTIC recently changed its position titles and descriptions related to the introduction of its EDMS equipment and workflow. The position previously related to micrographics (Xerographic Specialist) has been retitled Electronic Document Technician. This position performs all the functions required to create electronic images and archival quality silver halide master microfiche, to quality inspect those electronic images and microfiche, and to prepare diazo microfiche for delivery through DTIC's Automatic Document Delivery (ADD) subscription service. The incumbent serves as a systems administrator and trouble-shooter for the Branch by providing expertise in identifying, trouble-shooting and resolving a variety of system software and hardware operating problems. The grade and series was upgraded to a GS-303-6 based on the additional decisions required with the new equipment.

The optical character recognition is performed by a Data Verification Assistant. The position was previously titled Data Transcriber. The change in the title indicates a shift from data capture to data verification. The scanning and OCR is now performing the transcription function, converting the hardcopy to electronic. The emphasis for the operator is on the verification and correction of what the technology outputs. The grade and series have also been upgraded for this position to a GS-303-6 based on the increased responsibility and decision making regarding the quality of the output.

2.1.3 *National Library of Medicine*

NLM currently has no positions using these technologies. However, there is an RFCA being developed that includes these functions. The representatives were not at liberty to discuss the specifics of the RFCA. The specifications will be shared with the team when the procurement is completed.

2.1.4 *National Technical Information Service*

NTIS has not changed its position descriptions for the current staff performing scanning and image-based output for document delivery. The position titles remain Bindery Equipment Operator and Micrographics Clerk. The position grades and series remain XP-7 and GS-350. The XP is in the equipment operator series, while the GS is the general clerical series.

However, NTIS has on-site contractors performing contract work for other government agencies that includes scanning and OCR activities. These on-site contractors have modified their hiring practices and have changed the pay scale for these positions. NTIS recently created a DocuTech operator position description at a GS-5/6 for this outside contract.

Creating the product from the electronic version is getting so easy now at NTIS that the skill requirements are greatly reduced. The Xerox DocuPrint 6135 requires only loading of paper and taking the books out of the hopper when they are completed. The DocuPrint also has full binding capabilities. Future plans call for only one DocuTech (the predecessor to the DocuPrint) to be maintained. The DocuTech requires more operator intervention.

NTIS employees manning the DocuPrints/DocuTechs are XP-7's, positions with salary levels which may not be competitive with other Federal positions or with contract costs for performing similar work. Replacing incumbents lost through attrition with GS-5/6 level personnel or contracting out the work may be a necessary cost containment move for NTIS in the future.

While the technology is getting easier to use, the operators are making more decisions with the scanner than with the TDC camera, the technology being replaced. The pre-scan stage, where potentially difficult pages to scan are identified, requires keen judgement. The person performing the pre-scan is almost evaluating the importance of the document content.

NTIS currently is scanning reports only when they are requested for document delivery. A group is meeting at NTIS to discuss the scanning of newly received reports.

2.2 Other Government Agencies

Three non-CENDI government agencies were contacted regarding their scanner/OCR related position descriptions. Both the National Agricultural Library and the Library of Congress have digital library initiatives.

2.2.1 *National Agricultural Library*

The National Agricultural Library has not changed the title from Library Technician (Data Transcription), but the position description has been changed. The modification to the description includes the requirement to use a microcomputer and understanding of the scanner and OCR software. The grade/series remains a GS-1411-7, which is in the library technician series.

2.2.2 *Library of Congress, Digital Library Project*

The Library of Congress has five position descriptions for Digital Conversion Specialist, located in the Production Coordinating Group of the National Digital Library program. In consultation with the Library's custodial divisions, the Production Coordinating Group oversees the preparation and digitization of the historical collections. Activities include coordinating the organization, preservation and description of original archival materials; arranging for and monitoring the work of contractors who reproduce the items in digital form; and overseeing the placement of the delivered materials in a digital repository.

The GS-7/9/11/12 positions are in the GS 1001 series. In the developmental capacity of the GS-7, the Digital Conversion Specialists work in one or more specialized areas including SGML markup, multimedia elements, material selection and preliminary historical research, editorial assistance and computer technical support to the group. In the GS-9/11/12 positions, the Digital Conversion Specialist works in increasingly responsible capacities in these same areas.

The Senior Digital Conversion Specialist and Team Leader are in the GS-1410-13 part of the librarian series. The incumbents in these positions serve as experts in the specialized areas described above, with the exception of computer technical support.

2.2.3 *Congressional Research Service*

The Document Technician prepares documents for scanning to the optical disks, scans them, and quality reviews the scanned product before it is written to the disk. This individual also produces paper copies for materials listed on the order sheets for the weekly Selective Dissemination of Information (SDI) service or in the CRS Bibliographic Databases in SCORPIO. Kris Vajs (NTIS) formerly of the CRS, noted that because newer scanners require more than rote operation, the grade levels were changed from a 3/4/5 to a 5/6 following an upgrade of the scanners.

2.3 Private Organizations

The task group attempted to include information from relevant private sector organizations in this study. The first point of contact was the National Federation of Abstracting and Information Services (NFAIS), a major trade organization for database producers and other information companies. NFAIS last collected and published information concerning the position descriptions and pay scales within the database production industry in 1992. This survey did not include positions for scanner or OCR operators/technicians. No new information is available.

General searches on the WWW and specific searches of the Association for Information and Image Management (AIIM) home page were also performed. No position survey information was available free of charge. No information was provided from the Association for Workprocess Improvement (formerly, the Recognition Technologies Association), a trade association for data capture organizations.

Most private sector organizations contacted were hesitant to share position descriptions and compensation information. However, Chemical Abstracts Service (CAS) responded with a position description for a Desktop Graphics Specialist, part of the desktop publishing for the American Chemical Society (ACS) journals. The main responsibility is to deal with manuscripts provided by authors and to scan and enhance the graphics. The task group considered this position to be too different from the positions being analyzed within the CENDI agencies, so the CAS description was not used in the analysis and is not presented in the comparison chart.

3.0 ANALYSIS

Scanning and OCR technologies are used both as part of data capture in the front end of the workflow and at the point of document distribution. In recently implemented electronic document management systems, document distribution is performed by converting an image to another form of output, such as hardcopy or microfiche. In the case of the front-end data capture, the staff who currently use this technology were previously typists, data capture technicians, or proofreaders. The document distribution use of the technology is generally performed by someone who previously worked in micrographics or offset printing.

The technology for data capture is very integrated with the data capture process. It varies from agency to agency depending on the workflow and the data to be captured. While scanning and optical character recognition is likely to remain into the future, some of this effort (perhaps the majority) will be superseded by electronic receipt of documents.

The technology for document distribution is less integrated with the workflow process. It is meant to be performed on an on-demand basis, taking stored documents in electronic form and providing output at the request of the customer in a variety of media and formats. It also may involve the conversion to electronic form of documents that are in the legacy collection when they are requested. Electronic to hardcopy conversion will continue for some time into the future both for domestic and international recipients.

Upon evaluation of the descriptions that have been modified, it should be noted that the descriptions incorporate scanner and OCR technologies simply as additional tools to perform the work, similar to how the microcomputer was incorporated into the data capture process several years ago. The major emphasis of the position description is the work to be performed, not the technology used to perform it.

It should also be noted that position incumbency has been a driver during this transition period from hardcopy to electronic. Current staff, many of whom have been in the positions for many years, are not necessarily working under the appropriate grades and series. However, as staff turnover takes place, the need will arise to provide appropriate position descriptions.

The task group determined that the agencies are generally in a state of transition with regard to the incorporation of scanner and OCR technologies. In order to incorporate the new technologies gradually, agencies retrained existing staff to use them in revised workflows. The staff generally performed the data capture functions (most recently via microcomputer or mainframe terminals) at the beginning of the workflow, or they performed reprographics functions (blowback to hardcopy, microfiche production, or photocopying) at the end of the workflow for document distribution.

The employees in data capture, who may review the OCR output, were previously cataloging or data capture clerks, or technicians. Those performing the high speed scanning or blow-back from the image files were often retrained from among the offset press operators or reprographic staff.

Transitioning incumbents is an important key to successful implementation of electronic document management. No organization interviewed with the possible exception of the LoC National Digital Library project was able to contract or hire new people to perform these new functions. Budget constraints and in some cases hiring freezes as well as employment policies required that existing staff be used. (More detailed discussions about retraining are included in “Scanning and OCR Technologies Among the CENDI Agencies,” CENDI/96-1.)

A major concern for the agencies is the allocation of portions of the EDMS workflow to employees at appropriate levels with appropriate skills, including the retraining of incumbents as necessary, and developing completely new PDs only when the reengineering or replacement of a workflow justifies the effort. DTIC's EDMS has been in place for approximately five years and the technologies are well integrated into its workflow. The reengineering and replacement of the workflow matured to the point where the PDs could be rewritten appropriately. Other agencies are still in a pilot or transition phase between old and new workflows.

4.0 RECOMMENDATIONS

Although the task group debated the usefulness of drafting a suggested PD for a scanner/OCR equipment operator, our analysis indicated that developing language which may be inserted into a variety of position descriptions, already existing or newly created, would be more appropriate at this time. This “plug-in” language describes the major duties of the scanning/OCR function and states the core competencies (knowledge, skills and abilities) required.

The suggested language, presented in Section 4.1, identifies two primary requirements for the use of this technology: the ability to use microcomputer hardware and software and the ability to make judgements concerning potential problems in obtaining the best scanned image possible from the page being reviewed. An element of graphic artistry is required. Several organizations noted that time must be spent “training the eye” to identify what will scan well and what will not. The operator must decide which image density should be used, which images require enhancement, etc.

4.1 Suggested Scanning/OCR Position Description Language

, Major Duties

Assures that the document is identified with the correct document identifier. Checks each page for visual integrity - the letters are clear, the entire page scanned, no unusual lines appear, etc. -- and for completeness.

Adjusts scanner operation based on an assessment of document quality, i.e., legibility, gray scale requirements, paper quality, document type, and sizes.

Identifies problems with the equipment and reports them to the proper authority.

, Knowledge Required by the Position

Knowledge sufficient to operate optical disk scanning and printing technology including knowledge of all optical scanners to be used in order to determine the most appropriate one and assuring the document is correctly prepared for that scanner.

Knowledge of the scanning principles, techniques, and equipment to make judgements regarding the optimal scanning equipment adjustments for various document types; e.g., gray scale, two-sided, high-density, paper weight, etc.

Knowledge of microcomputers and the software used to perform the task.

Knowledge of equipment capabilities and the organization's standards for quality digitization.

A practical knowledge of data verification, review, and editing techniques.

, Complexity

The range of types of materials complicates the work of the incumbent. Projects require resourcefulness and ingenuity in developing, producing and ensuring the high quality of digitized products.

Ability to pay attention to detail.

Ability to perform repetitive tasks with accuracy.

Ability to identify the resolution and enhancements to be applied to an item when it is scanned in order to produce the highest quality image from the equipment available.

The incumbent must be constantly cognizant of the requirement to balance production and timeliness goals with product quality imperatives.

5.0 CONCLUSION

This report is an evaluation midway in the agencies' transition to an electronic environment. As the percentage of information acquisitions handled electronically increases, personnel requirements to support the new workflows should continue to be examined.

Appendix I

CENDI SURVEY OF POSITION DESCRIPTIONS/GRADE LEVELS/PAY SERIES FOR SCANNER OPERATORS AND OCR TECHNICIANS

CENDI Information Exchange Working Group,
Scanner/OCR Technicians Position Description Task Group
Rev 2. September 23, 1996

NOTE: Response is needed by Friday, Sept. 27, 1996.

Do you have a position description(s) specific to scanner operators?

If yes, what is the position title(s)?

(Please provide a copy of the position description(s) to the address indicated at the end of this survey.)

If yes, what was the position description(s) of the employees prior to the introduction of this technology?

If there is no position description specific to scanner operators, what is the position title(s) currently being used for scanner operators?

(Please provide a copy of that position description(s) to the address at the end of this questionnaire.)

What is/are the grade level and pay series for the description(s) provided?

Are you satisfied that the position description(s) your organization is currently using adequately describes the position performed by your scanner operators? If not, what are the areas of concern?

Do you have efforts underway to change the position descriptions/level/series for the scanner operators? Please describe.

Do you have a position description(s) specific to OCR technicians?

If yes, what is the position title(s)?

(Please provide a copy of the position description(s) to the address indicated at the end of this survey.)

If yes, what was the position description(s) of the employees prior to the introduction of this technology?

If there is no specific position description for OCR technicians, what is the position title(s) currently being used?

(Please provide a copy of that position description(s) to the address at the end of this questionnaire.)

What is/are the grade level and pay series for the description(s) provided?

Are you satisfied that the position description(s) your organization is currently using adequately describes the position performed by your OCR technicians? If not, what are the areas of concern?

Do you have efforts underway to change the position description(s)/level/series for the OCR technicians? Please describe.

Respond before Friday, September 27, 1996 to:

Gail Hodge
Senior Analyst, CENDI Secretariat
e-mail: Gailhodge@aol.com
Fax: 610/789-6769 (please call first)
Phone: 610/789-6769

Appendix II

Job Description Comparison Chart

CENDI SCANNER POSITION SURVEY RESULTS

December 6, 1996

Agency	Position Title	PD Avail?	Changed?	Grade/ Series	Appropriateness	Future Plans
DOE	Printing Spec.; Offset Press Oper.; Offset Photographer; Technical Information Clerk	No	No	Not Applicable	Only a concern if the volume of work increases and optimum scanning cannot be performed as it is currently being done. (1)	Would change only if percentage of work changes.
DTIC	Electronic Document Technician	Yes	Yes Previous: Xerographic Specialist	GS-303-06		
NLM	None					

NTIS	Bindery Equipment Operator;
	Micrographics Specialist;
	on-site contractors
		Not Applicable; not avail-able for on-site contrac-tors	No	XP-7
				GS-350
				on-site contractors in the 5/6 range
					On-site contractors have hired at a lower grade than incumbents.	May need to modify routine agency positions to bring them in line with the on-site contractors.

(1) When the work can no longer be performed optimally as add-on duties to the current position, OSTI will revisit the arrangement.

CENDI OCR POSITION SURVEY RESULTS
December 6, 1996

Agency	Position Title	PD Avail?	Changed?	Grade/ Series	Appropriateness	Future Plans
DOE	None; runs automatically					
DTIC	Data Verification Assistant	Yes	Yes Previous: Data Transcriber	GS-303-06		
NLM	None					
NTIS	Only beginning to discuss					

OTHER GOVERNMENT AGENCIES - SCANNER POSITION SURVEY RESULTS

December 6, 1996

Agency	Position Title	PD Avail?	Changed?	Grade/Series	Appropriateness	Future Plans
Congressional Research Service	Document Technician	Yes	Yes; upgraded based on increased decision making	GS-1411-5/6		
National Agricultural Library	Library Technician (Data Tran-scription)	Yes	Title, grade/series have not changed. Position description has changed.	GS-1411-7		
Library of Congress	Digital Conversion Specialist	Yes	New position	GS-1001-7/9/11/12/13	Covers full range from novice to supervisor	