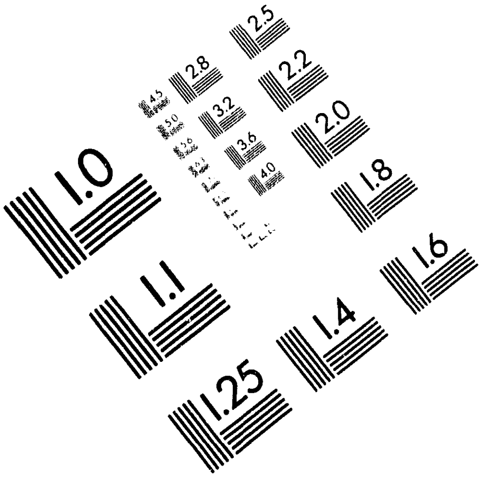
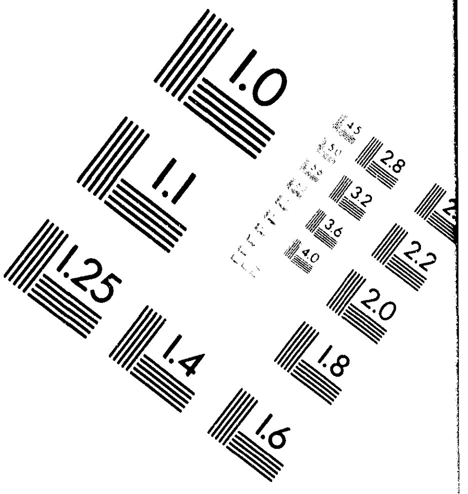


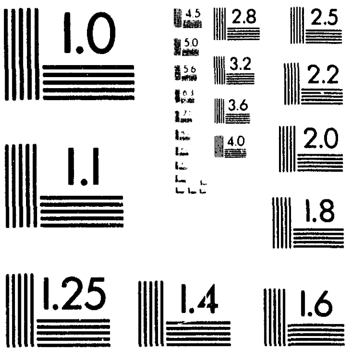
AIIM

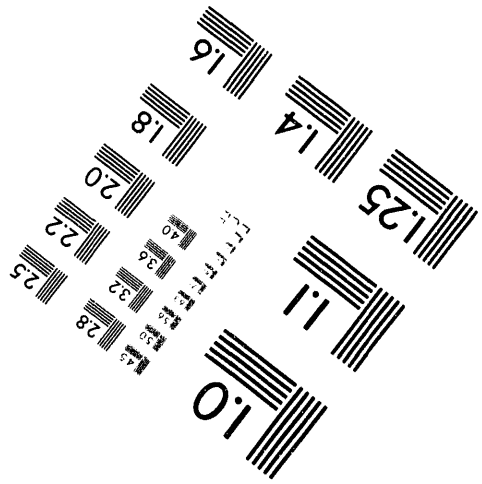
Association for Information and Image Management

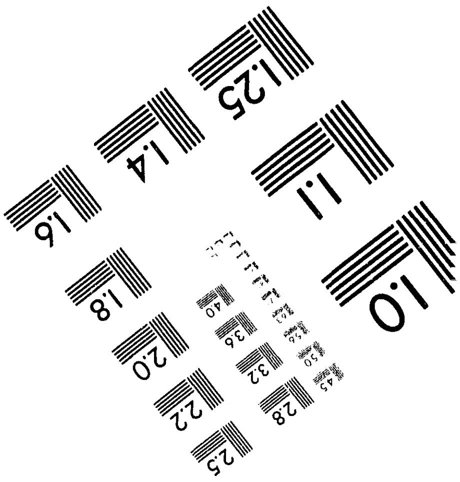
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202









MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

NOTICE

CERTAIN DATA
CONTAINED IN THIS
DOCUMENT MAY BE
DIFFICULT TO READ
IN MICROFICHE
PRODUCTS.

1 of 1

DOE/NV/10872--T105

UNLV
Information Science
Research Institute

Quarterly Progress Report

T. A. Nartker
March 31, 1994

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Table of Contents

UNLV Information Science Research Institute
Quarterly Progress Report

T. A. Nartker
March 31, 1994

I. Board of Advisors Activity

The ISRI Advisory Board will meet next during our 1994 Symposium.

II. Symposium Activity

The 1994 Symposium will be held on Monday, Tuesday, & Wednesday, April 11-13 at the Alexis Park Hotel in Las Vegas. Of the 60 papers submitted, 25 were accepted for presentation and 9 more were accepted as "poster" papers. Four invited talks will be delivered by Dr. C.K. Chow, Dr. G. Salton, Dr. G. Toussant, and Ms. D. Harman. The preliminary program was mailed in February (see Appendix A). Both the Proceedings and the 1994 Annual Report have been sent to the printer.

Currently we have 80 pre registrations which is about 40 more than in previous years at this time. Almost all preparations are complete.

III. Staff Activity

Recruiting

We are not recruiting at the current time.

Travel/ Meetings

In January, T. Nartker and K. Taghva traveled to San Diego, CA to attend a briefing at UCSD concerning the NSF "Digital Libraries" program.

In February, T. Nartker, K. Taghva, and F. Jenkins traveled to San Jose, CA to attend and present papers (3) at the annual SPIE Conference.

In March, S. Rice and L. Gewali traveled to Boca Raton to attend a Conference and present a paper at Florida Atlantic University.

Papers accepted or presented

A paper by K. Taghva, J. Borsack, A. Condit, and S. Erva on the "Effects of Noisy Data on Text Retrieval" was published in the January issue of JASIS.

T. Nartker delivered a keynote address, "On the Need for Information Metrics," before the 1994 Symposium on Electronic Imaging Science & Technology in San Jose, CA. in February.

A paper by J. Kanai and F. Jenkins titled, "Use of Synthesized Images to Evaluate the Performance of OCR Devices and Algorithms," was presented at the 1994 Symposium on Electronic Imaging Science & Technology in San Jose, CA. in February.

A paper by K. Taghva, J. Borsack, and A. Condit titled, "An Expert System for Automatically Correcting OCR Output," was presented at the 1994 Symposium on Electronic Imaging Science & Technology in San Jose, CA. in February.

A paper by K. Taghva, J. Borsack, B. Croft, and Steve Harding titled, "Evaluation of Retrieval Effectiveness of Simulated OCR Text," was accepted for presentation at the Third Annual Symposium on Document Analysis and Information Retrieval.

A paper by K. Taghva, J. Borsack, and A. Condit titled, "Results of Applying Probabilistic IR to OCR Text," was accepted for presentation at SIGIR94.

IV. Document Analysis Program

OCR Devices

Six companies have submitted OCR devices for the 1994 technology assessment tests. One company (TRW) has submitted a new system based on a "majority-voting" algorithm. This is the first "voting" based system we have tested except, of course, for the ISRI developed voting system.

OCR Test system

Version #4 of our experimental system will be used for our 1994 tests.

OCR Databases/GT1

Both the magazine sample and DOE sample #2 (used last year) will be used for our 1994 tests.

OCR Databases/Foreign languages

We have prepared ground-truth text for a set of 57 pages from printed Chinese documents. A preliminary report on Chinese OCR systems will be included in our 1994 Annual Report.

OCR Experiments

Our 1994 Technology Assessment tests have been completed. As was true last year, the report of these tests is part of our 1994 Annual Research Report which will be first distributed to attendees at the SDAIR Symposium.

OCR Technical reports/thesis

None

Interaction with OCR vendors

In January, John Martinez, president of CTA Corporation visited our lab for a demonstration. CTA is an OCR vendor with home offices in Barcelona, Spain.

We have been in close contact with Phil Cheatle of Hewlett Packard Laboratories in Bristol England. We expect they are a good candidate to join our Industrial Affiliates program.

Interaction with OCR research organizations

None

V. Text-Retrieval Program

TR Databases

We have completed minimum document verification (MDV) for approximately 900 of the usable documents in GT1. There are about 400 more documents which can be recovered. We are also collecting information such as relevancy judgments, concept information, keywords and key phrases.

TR Experiments/Projects

Our project to design and implement a new user interface to enable interaction between images and text has made significant progress. We expect to have a usable system in June.

We are beginning to examine the effect of SGML tagging on retrieval efficiency.

TR Technical reports/thesis

None

Document Routing Project

None

VI. Institute Activity

Institute visitors

Date	Visitor	Agency
01/12/94	John Martinez	CTA
01/13/94	B. Courtney, J. Arcos, & M. Smith	TRW
01/21/94	S. Dennis & R. Wenzel	DoD
02/18/94	M. Buchman & D. Hurry	DoD
03/05/94	Dr. Abraham Kandel	USF
03/21/94	Dr. Don Morrison	UNM

Institute seminars

"Intelligent Hybrid Systems," Professor Abraham Kandel, University of South Florida.

New agency contacts/ new research proposals

Current work on our Ft. Meade contract is focused on demonstrating our tools for testing foreign language OCR systems (i.e., with wide characters) via preliminary testing of Chinese OCR systems. We have also begun preparing Japanese "ground-truth" test data.

T. Nartker and K. Taghva have continued to discuss possible cooperative research with Los Alamos National Labs.

T. Nartker, K. Taghva, and J. Kanai have submitted a four year/multi-million dollar proposal to the "Digital Libraries" research program of NSF in cooperation with SRI International, the Desert Research Institute, and the Univ. of Nevada, Reno (see Appendix B).

VII. Goals Achieved/Goals for Next Quarter

Goals from last quarter:

1) We have begun preparing Sample#3 from the DOE documents. We are also beginning work on a "business letter" data set for 1995 tests.
2) Preliminary programs for SDAIR94 were mailed in February.
3) We have been unable to locate software support for managing unicode text files.
4) The third meeting of the Industrial Affiliate's program will take place during SDAIR94. No new members were added during the winter quarter.
5) The new user interface for text retrieval systems will be tested during the Spring quarter.
6) About 100 additional documents have passed MDV.

Goals for next quarter:

1) Conduct SDAIR94.
2) Conduct the third meeting of the members of the Affiliate's program. We will continue to recruit new members for this program.
3) Continue work on Sample#3 and a business letter test dataset.
4) Test the new user interface for text retrieval systems.
5) Continue to search for software to manage unicode text files.
6) Continue MDV for the remainder of GT1.
7) Prepare a plan for the 1995 technology assessment test program.

APPENDIX A.

Preliminary Program for SDAIR94

Third Annual Symposium on Document Analysis and Information Retrieval

April 11 - 13, 1994

Alexis Park Hotel
Las Vegas, Nevada

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Sponsored by the

Information Science Research Institute

and

The Howard R. Hughes College of Engineering

University of Nevada, Las Vegas

CONFERENCE SCHEDULE

Sunday, April 10, 1994

7:00pm - 10:00pm Alexis Park
Reception

Monday, April 11, 1994

7:00am - 8:20am Alexis Park
Registration

8:20am - 8:30am Alexis Park
Welcome

Theo Pavlidis, Chairman
Leading Professor
Department of Computer Science
State University of New York at Stony Brook

Robert C. Maxson, President
University of Nevada, Las Vegas

William R. Wells, Dean
Howard R. Hughes College of Engineering
University of Nevada, Las Vegas

8:30am - 9:15am Alexis Park
Invited Speaker

Recognition Error and Reject Trade-off
C. K. Chow, IBM Research Center (Emeritus)

9:15am - 9:30am Alexis Park
Refreshment Break

9:30am - 10:30am Alexis Park
Session 1 Chair: Jonathan Hull

Adaptive Logic Networks for Machine-Printed Character Recognition
Robert C. Vogt, John J. LoPorto, John M. Trenkle, William Cavnar, Environmental Research Institute of Michigan

A Comparison of Two Learning Algorithms for Text Categorization
David D. Lewis, AT&T Bell Laboratories;
Marc Ringuette, Carnegie Mellon University

Learning the Optimal Parameters in a Ranked Retrieval System Using Multi-Query Relevance Feedback
Brian T. Bartell, Garrison W. Cottrell, Richard K. Belew, University of California at San Diego

Pattern Classification Based on Adaptive Self-Organized Neural Network
Yuan-Yan Tang, Ching Y. Suen, Tao Li, Concordia University; L. Y. Fang, Bell-Northern Research

10:30am - 10:45am Alexis Park
Poster Break

10:45am - 11:30am Alexis Park
Session 2 Chair: Robert Korfhage

An Evaluation of Information Retrieval Accuracy With Simulated OCR Output
W. B. Croft and S. Harding, University of Massachusetts-Amherst; K. Taghva and J. Borsack, University of Nevada, Las Vegas

Validation of Simulated OCR Data Sets
George Nagy, Rensselaer Polytechnic Institute

Validation of Page Defect Models for Optical Character Recognition
Yanhong Li, Daniel Lopresti, Andrew Tomkins, Matsushita Information Technology Laborotory

11:30am - 11:45am Alexis Park
Poster Break

11:45am - 12:55pm
Lunch (no host)

12:55pm - 1:00pm Alexis Park
Announcements

1:00pm - 1:45pm Alexis Park
Invited Speaker

Text Retrieval Using the Vector Processing Model
Gerard Salton, Cornell University

1:45pm - 2:00pm Alexis Park
Break

2:00pm - 3:00pm Alexis Park
Session 3 Chair:

Keyword Selection From Word Recognition Results Using Definitional Overlap
Paul Filipski, Jonathan Hull, State University of New York at Buffalo

N-Gram-Based Text Categorization
William B. Cavnar, John M. Trenkle, Environmental Research Institute of Michigan

Lexicon-Based Word Recognition Without Word Segmentation
Gregory K. Myers and Chien-Huei Chen, SRI International

2:00pm - 3:00pm Alexis Park
Session 3 (Continued)

An Automatic Indexing of Compound Words Based on Mutual Information for Korean Text Retrieval
Pan Koo Kim, Yoo Kun Cho, Seoul National University

3:00pm - 3:15pm Alexis Park
Poster Break

3:15pm - 4:00pm Alexis Park
Session 4 Chair: David Lewis

The Role of Visualization in Document Analysis
Robert R. Korfhage, University of Pittsburgh; Kai A. Olsen, Molde College, Norway

About the Logical Partitioning of Document Images
Andreas Dengel, German Research Center for Artificial Intelligence

A Context-Based Approach to Text Recognition
T. G. Rose, L. J. Evett, Amanda Caryn Jobbins, Nottingham Trent University, Nottingham, England

4:00pm - 4:15pm Alexis Park
Poster Break

5:45pm - 11:00pm
Buses to Lake Mead Lake Mead
Dinner/Dance Cruise
on the Desert Princess

Tuesday, April 12, 1994

7:30am - 8:20am Alexis Park
Registration

8:20am - 8:30am Alexis Park
Welcome

Theo Pavlidis, Chairman
Leading Professor
Department of Computer Science
State University of New York at Stony Brook

8:30am - 9:45am Alexis Park
Invited Speaker

Computational Geometry for Document Analysis
Godfried Toussaint, McGill University, Montreal, Canada

9:45am - 10:00am Alexis Park
Refreshment Break

10:00am - 11:00am Alexis Park
Session 5 Chair:Henry Baird

Script and Language Determination From Document Images
A. Lawrence Spitz, Fuji Xerox Palo Alto Laboratory

Binarization and Multi-Thresholding of Document Images Using Connectivity
Lawrence O'Gorman, AT& T Bell Laboratories

Direct Extraction of Topographic Features From Gray Scale Character Images
Seong-Whan Lee and Young Joon Kim, Chungbuk National University, Korea

An Alternative to Vectorization: Decomposition of Graphics Into Primitives
J. E. den Hartog, T. K. ten Kate, G. van Antwerpen, TNO Institute of Applied Physics; J. J. Gerbrands, Delft University of Technology; The Netherlands

11:00am - 11:15am Alexis Park
Poster Break

11:15am - 12:00pm Alexis Park
Session 6 Chair: Larry Spitz

Asymptotic Accuracy of Two-Class Discrimination
Tin Kam Ho and Henry S. Baird, AT&T Bell Laboratories

Use of Constraints As A Second Stage Character Classification Technique
George Sazaklis and Theo Pavlidis, State University of New York at Stony Brook

Performance Evaluation of Two OCR Systems
Su Chen, Suresh Subramaniam, Robert M. Haralick, University of Washington; Ihsin T. Phillips, Seattle University

12:00pm - 1:10pm
Lunch (no host)

1:10pm - 1:15pm Alexis Park
Announcements

1:15pm - 2:00pm Alexis Park
Invited Speaker

The Text REtrieval Conference
Donna Harman, National Institute of Standards and Technology

2:00pm - 2:15pm Alexis Park
Break

2:15pm - 3:15pm Alexis Park
Session 7 Chair: Robert Korfhage

An Approach to Interactive Retrieval in Face Image Databases Based on Semantic Attributes
Venkat N. Gudivada, Ohio University; Vijay V. Raghavan, Guna S. Seetharaman, University of Southwestern Louisiana

Marking of Document Images With Codewords to Deter Illicit Dissemination
J. T. Brassil, S. Low, N. F. Maxemchuk, L. O'Gorman, AT&T Bell Laboratories

Modelling and Exploiting Traceability Between Software Development Documents
Jean-Pierre Queille, Anne Richermo, Jean-Francois Voidrot, Matra Marconi Space; Florence Sedes, University of Paul Sabatier

Development of a Full-Text Information Retrieval System
Keizo Oyama, Akira Miyazawa, Atsuhiro Takasu, National Center for Science Information Systems (NACSIS); Kouji Shibano, Tokyo International University

3:15pm - 3:30pm Alexis Park
Poster Break

3:30pm - 4:30pm Alexis Park
Session 8 **Poster Papers**

Low Level Structural Recognition of Documents
Abdel Belaid, Y. Chenevoy, CRIN-CNRS/INRIA Lorraine, France

Document Characterization, Authentication and Retrieval Based on Medium-Embedded Random Patterns
D. Brzakovic and N. Vujovic, Lehigh University

Information Retrieval for a Document Writing Assistance System
Marie-Louise Corral and Amaury Simon, Matra Marconi Space (MMS) - Aramiihs; Christine Julien, Institut de Recherche en Informatique de Toulouse (IRIT); France

Issues in Automatic OCR Classification
Jeffrey Esakov, Daniel P. Lopresti, Jonathan S. Sandberg, Jiangying Zhou, Matsushita Information Technology Laboratory - Panasonic Technologies, Inc.

Correlated Run Length Algorithm (CURL) for Detecting Form Structure Within Digitized Documents
Michael D. Garris, National Institute of Standards and Technology

Experiments in Automatic Word Class and Word Sense Idenification for Information Retreival
Susan Gauch, University of Kansas; Robert P. Futrelle, Northeastern University

Estimating Errors in Document Databases
Jaekyu Ha, Su Chen, Robert M. Haralick, University of Washington; Ihsin T. Phillips, Seattle University

Retrieval of Line Drawings
Oliver Lorenz and Gladys Monagan, Swiss Federal Institute of Technology (ETH)

A Paper Form Processing System With an Error Correcting Function for Reading Handwritten Kanji Strings
Katsumi Marukawa, Kazuki Nakashima, Masashi Koga, Yoshihiro Shima, Hiromichi Fujisawa, Central Research Laboratory, Hitachi, Ltd.

5:30pm - 10:00pm
Happy Hour Great Hall
Dinner Thomas Beam
Tour of Facilities Engineering Bldg.
UNLV

Wednesday, April 13, 1994

8:20am - 8:30am Alexis Park
ISRI Welcome

Thomas A. Nartker, Director
Information Science Research Institute
Howard R. Hughes College of Engineering
University of Nevada, Las Vegas

8:30am - 9:45am Alexis Park
The 1994 ISRI Technology Assessment Reports
ISRI Staff

9:45am - 10:00am Alexis Park
Refreshment Break

10:00am - 12:00am Alexis Park

ISRI Research Reviews
ISRI Staff

Invited Speakers

C. K. Chow is a Research Staff Member Emeritus of IBM. He joined IBM in 1964 at the Thomas J. Watson Research Center where he did research and had various managerial, staff and international assignments. Prior to that he worked at the Burroughs Research Center. He received his Ph.D. degree from Cornell University and served as an assistant professor at Penn State, a visiting professor at MIT, an adjunct professor at Columbia and UC Santa Barbara. His professional experience included membership on the IEEE Computer Society Governing Board, conference chairmanship and associate editorship of several journals. He is a life fellow of the Institute of Electrical and Electronic Engineers.

Gerard Salton is a Professor of Computer Science at Cornell University. He received a Ph.D. degree in Applied Mathematics from Harvard University in 1958 and remained on the Applied Mathematics faculty at Harvard until 1965 when he moved to Cornell. Dr. Salton has been a visiting professor at the University of Grenoble (France), at the Swiss Federal Institute of Technology (ETH) in Zurich and (EPF) in Lausanne and at the University of Konstanz (Germany). Dr. Salton has worked in the area of natural language text processing, including, in particular, automatic text analysis and information retrieval. From 1965-68, he was editor-in-chief of ACM Communications; from 1969-70, editor-in-chief of the ACM Journal. Between 1972 and 1978, he served on the ACM Council as Northeast Regional Representative. Currently, Dr. Salton is an editor of Information Systems and of the ACM Transactions on Database Systems. He was a Guggenheim Fellow in 1963 and has received the first ACM-SIGIR Award for contributions to information retrieval in 1983, as well as a Humboldt Foundation Senior Scientist Award in 1988 and the ASIS Award of Merit in 1989. He has published a large number of articles and several books on information retrieval and related areas. The most recent text is "Automatic Text Processing" (Addison-Wesley, 1989).

Godfried T. Toussaint received his B.Sc. degree from the University of Tulsa, Tulsa, Oklahoma and his M.A.Sc. and Ph.D. degrees from the University of British Columbia, Vancouver, B.C., Canada, in 1968, 1970 and 1972, respectively, all in Electrical Engineering. Since 1972 he has been with the School of Computer Science at McGill University teaching and doing research in the areas of information theory, pattern recognition, and computational geometry. Dr. Toussaint is past council-member of the North American Branch of the Classification Society and past Associate Editor of the IEEE Transactions on Information Theory and of the IEEE Transactions on Pattern Analysis and Machine Intelligence. Presently, he is Associate Editor of the Plenum Press Series on Advanced Applications in Pattern Recognition, Associate Editor of Pattern Recognition, Associate Editor of Computational Geometry: Theory and Applications, Associate Editor of the International Journal of Computational Geometry and Applications and Associate Editor of the Visual Computer. He is also on the Editorial Boards of the Journals Discrete and Computatational Geometry and Forma as well as on the Advisory Board of the IEEE Transactions on Pattern Analysis and Machine Intelligence. He is a member of several learned societies including the IEEE, The Pattern Recognition Society and the New York Academy of Sciences. He recently edited two books published by North Holland, Computational Geometry in 1985 and Computational Morphology in 1988 as well as three special issues on computational geometry, one of The Visual Computer (May, 1988), one of the Proceedings of the IEEE (September, 1992), and one of Pattern Recognition Letters (September, 1993). In 1978, he was the recipient of the Pattern Recognition Society's Best Paper of the Year Award and in 1985 he was awarded a Killam Fellowship by the Canada Council to carry out a two-year research project on movable separability of sets.

Donna Harman has been involved in research in new retrieval techniques for many years. She works at the National Institute of Standards and Technology (NIST) and has built a large-scale prototype of an advanced retrieval system for testing in several government agencies. Currently, she is involved in running the Text Retrieval Conferences (TREC), including developing a new test collection involving over a million documents, with appropriate topics and relevance judgements. She received an M.E.E. degree in electrical engineering from Cornell University, and worked with Professor Gerard Salton on the SMART project. Before coming to NIST, she was a researcher at the National Library of Medicine in the areas of expert systems in medicine and informational retrieval systems.

Third Annual Symposium on Document Analysis and Information Retrieval

INFORMATION SCIENCE RESEARCH INSTITUTE
University of Nevada, Las Vegas
April 11-13, 1994

Conference Registration Form

Name: ______________________________

Title: ______________________________

Company: ______________________________

Address: ______________________________

City: ______________ State: __________ Zip: __________

Telephone Number: (   ) ______________________________

E-mail Address: ______________________________

Registration Fees	Pre-Reg before 3/11/94	Regular after 3/11/94	Amount
Conference Registration (Includes dinner Tuesday, 4/12/94)	$375.00	$450.00	$________
Dinner (Tuesday Dinner for Spouse/Companion)		$ 10.00	$________
Dinner/Cruise on Lake Mead (Monday Dinner) (Pre-registration is recommended since seating is limited)		$ 50.00	$________
Conference Proceedings (Extra Proceedings) (One Proceedings is included in the price of the registration)		$ 50.00	$________

Make checks/money orders payable to: **UNLV Board of Regents**

Mail completed conference registration form and check/money order to:
Symposium Manager
Information Science Research Institute
University of Nevada, Las Vegas
4505 Maryland Parkway
Box 454021
Las Vegas, NV 89154-4021

Telephone	(702)895-4571
Fax	(702)895-1183

All checks/money orders should be in U. S. Dollars and checks must be drawn on a U.S. Bank.

HOTEL REGISTRATION FORM

Rooms Reserved Under the Name: Third Annual Symposium on Documentation

Reservations received after **March 10, 1994** will be accepted on a space available basis only.

Please reserve accommodations for:

NAME: ______________________________

HOME ADDRESS: ______________________________

CITY: ______________ STATE: ______________ ZIP: __________

HOME PHONE: ______________________________

COMPANY NAME: ______________________________

COMPANY ADDRESS: ______________________________

CITY: ______________ STATE: ______________ ZIP: __________

BUSINESS PHONE: ______________________________

SINGLE OCCUPANCY - $ 79.00 DOUBLE OCCUPANCY - $ 79.00

WILL ARRIVE: ______________, 1994 TIME: ______________

WILL DEPART: ______________, 1994 TIME: ______________

Reservations will not be held after 30 days without a deposit. Credit Card Numbers are taken as a guarantee only, not as a method of payment.

Enclosed is my one night's deposit payable by: Check **(Circle One)** Credit Card

Mastercard _____ Visa _____ American Express _____

Carte Blanche _____ Diners Club _____

CREDIT CARD NUMBER: ______________________________

EXPIRATION DATE: ______________________________

PLEASE PRINT NAME AS IT APPEARS ON CARD: ______________________________

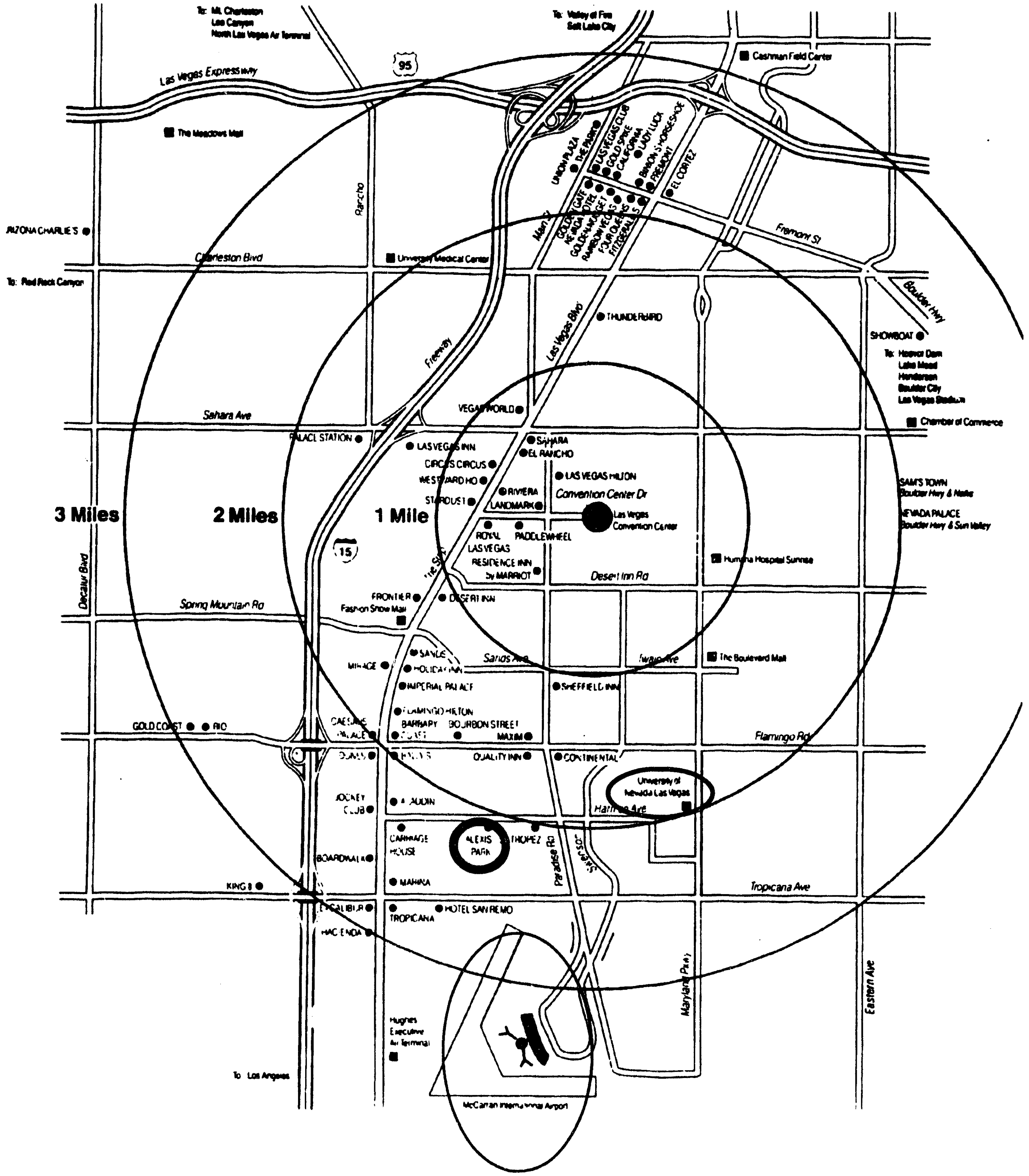
Room Reservations: (800) 582-2228 **Fax: (702)796-4334**

Or mail your reservation to:

Alexis Park Resort
P.O. Box 95698
Las Vegas, NV 89193-5698

Las Vegas

Rooms & Accommodations



BIOGRAPHIES

M. Bienkowski, SRI International
K. Finn, SRI International
L. Gewali, UNLV/ISRI
J. Hastings, The Desert Research Institute
J. Hobbs, SRI International
J. Kanai, UNLV/ISRI
S. Latifi, UNLV/ISRI
L. Larmore, UNLV/CSD
P. Mulgaonkar, SRI International
G. Nagy, Rennselaer Polytechnic Institute
T. Nartker, UNLV/ISRI
P. Stubberud, UNLV/ECE
K. Taghva, UNLV/ISRI
R. Wharton, The Desert Research Institute
S. Zink, University of Nevada, Reno

APPENDIX B.

The Automated Recovery and Organization
of Scientific and Technical Information
from Microform Archieval Documents

A Proposal
submitted to:
the NSF/ARPA/NASA Digital Libraries Program

DIGITAL LIBRARIES: THE AUTOMATED RECOVERY AND ORGANIZATION OF SCIENTIFIC AND TECHNICAL INFORMATION FROM MICROFORM ARCHIVAL DOCUMENTS

Submitted To:

The RESEARCH ON DIGITAL LIBRARIES Program

Prepared By:

T. A. Nartker, K. Taghva, & J. Kanai
The Information Science Research Institute
University of Nevada, Las Vegas
Las Vegas, Nevada

P. Mulgaonkar & J. Hobbs
SRI International
Menlo Park, California

R. Wharton & J. Hastings
The Desert Research Institute
Reno, Nevada

S. Zink
University of Nevada, Reno
Reno, Nevada

February 1994

TABLE OF CONTENTS

PROJECT SUMMARY

The objectives of this research are to identify and develop the technologies needed to support end-to-end, automated recovery and organization of scientific and technical information from microform documents. Historically, a large body of existing information has been produced on hard copy and archived on microform. Although current document processing systems are capable of converting clean printed documents into digitized form, recognition of text from poor quality documents is error prone. Furthermore, these systems cannot automatically organize extracted information. Therefore, information recovery from microform documents is an appropriate research area.

In this project, research problems in the following four stages of document conversion are investigated: pre-processing, recognition of document objects, retrieval of information, and browsing and discovery. Preprocessing steps serve two roles: 1) they remove scanning artifacts from images prior to recognition, thereby improving recognition rates; 2) they partition the image into regions based on information modality so that each region can be processed by suitable recognition systems. In particular, correction of geometric distortion, removal of photometric noise, and page decomposition problems will be investigated.

Currently, text regions are converted into ASCII using optical character recognition (OCR) systems Since errors generated by OCR systems propagate to down stream applications they strongly affect the overall performance of information processing systems. Thus, text-based methods for correcting OCR errors will be studied.

Graphic regions are usually manually indexed. Thus, information that does not correspond to a predefined a set of keywords can be neither extracted nor accessed. To overcome these problems, a model-based approach to automated extraction of information and methods for automated linking of text and graphics will be examined.

Although the text-based correction methods are expected to correct many OCR errors, it is highly likely that some character errors will remain in OCR generated text. Hence, the effect of OCR translation errors on information retrieval must be investigated. Furthermore, to deal with OCR errors, approximate matching techniques will investigated.

To make extracted information useful, methods to automatically organize the information will be explored. Some of the proposed research topics include: 1) automatic markup of the structure of a document, 2) automated indexing of documents, 3) natural language-based query construction, 4) natural language-based precision refinement, and 5) automated generation of hypertext links for browsing,

Progress made in this project will be measured using prototypes of an end-to-end data recovery system, a set of real world documents, and a set of real world queries. In particular, technical papers in the NSF Antarctic literature database will be used in the evaluation process. Knowledge obtained from this project will be applicable to conversion of other corpora of printed/microform documents.

EXECUTIVE SUMMARY

The goal of the NSF/ARPA/NASA Digital Libraries Initiative is to achieve an economically feasible capability to digitize massive corpora of information from heterogeneous sources such that they can be stored, managed, searched, and retrieved automatically. To this end, the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas (UNLV), SRI International (SRI), the Desert Research Institute (DRI), and the University of Nevada, Reno (UNR), propose to undertake a cooperative project of applied research to identify and develop the technologies needed to support end-to-end, automated recovery and organization of scientific and technical information from microform archival English language documents.

We know that researchers do not read documents in a linear fashion and that many users prefer browsing through information, jumping from topic to related topic. Hypertext systems permit such information access by providing links between related portions of a document, and between related information in different documents. Hypertext authoring systems allow information creators to produce such links while the information is being generated.

However, there exists a large body of existing information that has historically been produced on hardcopy. Formats, document structures, and styles followed in these documents have been driven by visual aesthetics or by the linear form dictated by the medium. Document processing systems today are capable of converting much of this information from hardcopy into digital form.

Current document understanding systems cannot organize the converted information into the new structures required by users of digital information. Further, recognition of text from poor quality (old) documents is error prone, and the resulting character errors can affect the recall and precision of the information retrieval processes that operate on the recognized results.

We propose a focused research program that addresses these issues. Specifically, we propose:

- Selecting a well defined corpus: the 40,000 microform English language documents in the NSF Antarctic research database that covers all known publications in diverse but related scientific fields (geology, biology, atmospheric science, etc.) on the continent.
- Working with researchers who use this information now and will be the beneficiaries of the digital corpus when it is created, to identify the information access mechanisms that will be most useful to them
- Developing and characterizing document processing techniques that scan, preprocess, recognize, and organize the information extracted from the microforms based on the inputs from the user community.
- Developing a testbed with progressively increasing capabilities that: (a) the users can use, evaluate, and critique; (b) that can support the needs of the research team by providing accurate and representative data for evaluating and extending processing algorithms; and (c) that can serve as a testbed for integrating new research ideas and algorithms in a controlled and cooperative manner.

At the end of the four year research program, we expect to have added new and significant knowledge to the Information Storage and Retrieval literature. Specifically, techniques using natural language for extracting and linking information in documents; cataloging techniques that can be tailored by users to present specific views of the retrieved information; and techniques for adapting to and correcting errors caused by incorrect recognition.

Our research will also add to the literature in the document understanding arena, making fundamental progress in modeling noise and degradation in microform images; trainable domain-specific algorithms for page decomposition; intelligent voting techniques for combining the results of multiple recognition modules; graphic analysis techniques; and techniques for automatically selecting the best combination of algorithms based on measured document image quality. We will also publish results of detailed empirical characterizations of existing and new algorithms in all aspects of the conversion process, including preprocessing, recognition, precision, and recall. Performance of browsing systems (although necessarily subjective) will also be published for the research community to evaluate.

The algorithms and prototype systems that we develop will be made available to several communities: First, the digital library research community will benefit by having the large corpus available for experiments in automatic indexing, information organization, retrieval, and similar studies. Second, the document understanding community will benefit from the large volume of accurately truthed images that will be available to test new algorithms and systems. The lack of sufficient data, especially in the science and technology fields is often cited as a key restraint on the development of improved document understanding systems. Third, the researchers who study the polar Antarctic region, will benefit by having a complete compendium of knowledge in their research field available at their fingertips. It would be interesting to study how their research habits change once this source becomes readily accepted. Finally, commercial information providers will benefit by virtue of having a system that can be used to automatically (hence cost effectively) convert, organize and sell other corpora of documents that currently exist only in hardcopy form. We propose providing free access to the first three communities and licensing the technology to the commercial information providers.

Our research plan is focused on the technology voids that we have identified. The plan is subdivided into four elements: preprocessing technology, recognition technology, postprocessing technology, and use of natural language for browsing and cataloging. This research plan is backed by a well defined experimental methodology for characterization of algorithms and measurement of performance.

Over the past three years, the experimental environment required to compare, evaluate, and experiment with competing technologies in both optical character recognition (OCR) and information retrieval (IR) has been developed and used at ISRI in Las Vegas. The operation of such large scale test facilities requires a significant investment in both equipment and personnel. ISRI has developed a management infrastructure to conduct annual technology assessment tests of competing OCR systems. New performance metrics, new test datasets, and facilities to create new test datasets have been developed. The mechanism of publishing the results of this competition is a Symposium, "the Symposium on Document Analysis and Information Retrieval," held each year in Las Vegas.

Our team is strong and uniquely capable of conducting the proposed research. The Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas was established in 1990 by a grant from the United States Department of Energy. The overall mission of ISRI is to foster the improvement of automated technologies for document processing. Since 1990, ISRI has been conducting large scale experiments in document understanding using its automated experimental environment. No other research institute in the world currently has such capabilities. We have a track record in characterization of OCR technology and in investigation of the interaction between the recognition and retrieval tasks.

SRI International is the world's largest independent nonprofit research institute, chartered in the state of California, performing a broad spectrum of problem-oriented research. Groups at SRI that are involved in the proposed work have a significant track record in OCR and image understanding (e.g., for the US Postal Service), hypertext and multimedia, artificial intelligence, and distributed information storage and retrieval. Our natural language system (FASTUS), developed in 1992, was among the top two or three leaders in two recent evaluations of written language systems, and an order of magnitude faster than comparable systems.

The Desert Research Institute is the world's largest multidisciplinary organization conducting environmental research in arid lands. DRI was recently awarded a 6-year grant to study the McMurdo Dry Valleys region of Antarctica - a cold desert ecosystem - as part of the NSF's Long-Term Ecological Research (LTER) program. Over 25 senior investigators and graduate students are involved in this LTER project, making DRI home to the majority of active researchers in this area. Project staff are intimately familiar with the Antarctic literature regarding the Dry Valleys: they have studied all of it, and developed a great deal of it: they are the domain experts.

DRI originally grew out of the University of Nevada, Reno; and the two institutions maintain a close working relationship today, not only in science, but in service facilities as well, including their Libraries. For a number of years, the UNR Library has aggressively pursued electronic storage and retrieval for its holdings, which include a large collection of Government documents, on environmental matters in particular. Library staff are professionally interested in the effects of information technology on scientists' research habits and scientific productivity.

Under this program, ISRI will conduct research in OCR techniques, information retrieval, and document markup. ISRI will also be responsible for the generation of image and truthed character data for supporting the team's research activities. In cooperation with the other team members, ISRI will design, develop, document, operate, and maintain the experimental testbed. SRI will focus on preprocessing algorithms, graphics understanding techniques, and natural language techniques based on SRI's FASTUS system. DRI will provide the team access to the documents in the Polar Antarctic database and access to researchers who use the information. DRI scientists will also serve as test users of the data as it is converted and throughout the development cycle, work closely with the team in defining needs and evaluating technical approaches to meet the needs. UNR will provide bibliographical knowledge that will be required to organize the extracted information according to accepted conventions and patterns. UNR will also have the responsibility to obtain all required copyright releases and other administrative protocols for providing users controlled access to the data.

The entire team is committed to achieving the proposed goals and is enthusiastic about the program potential. Both the administration at UNLV and UNR as well as the management at SRI and DRI strongly support the proposed project.

In summary, we confidently look forward to achieving the proposed goals. The end results of our research will benefit the entire infrastructure of information extraction and organization on which the future of digital libraries will be based.

DIGITAL LIBRARIES:
THE AUTOMATED RECOVERY AND ORGANIZATION OF SCIENTIFIC AND TECHNICAL INFORMATION FROM MICROFORM ARCHIVAL DOCUMENTS

T. A. Nartker, K. Taghva, and J. Kanai
The Information Science Research Institute
University of Nevada, Las Vegas
Las Vegas, Nevada

P. Mulgaonkar & J. Hobbs
SRI International
Menlo Park, California

1. INTRODUCTION

One goal of the NSF/ARPA/NASA Digital Libraries Initiative is to achieve an economically feasible capability to digitize massive corpora of information from heterogeneous sources such that they can be stored, managed, searched, and retrievad automatically. To this end, the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas (UNLV), SRI International (SRI), the Desert Research Institute (DRI), and the University of Nevada, Reno (UNR), propose to undertake a cooperative project of applied research to identify and develop the technologies needed to support end-to-end automated recovery and organization of scientific and technical information from microform archival English language documents.

The large amount of extant archival information that is available only in microform makes this a particularly interesting problem domain from the digital libraries perspective. Unlike new information that can be organized from its creation to take advantage of the search capabilities of digital media, microform documents are reflections of the organization imposed by their hardcopy originals. We propose to study the methods by which end-users access such archival information and to reflect these needs in the document processing steps. This will allow us to automatically organize extracted information in ways that provide user friendly search capabilities which significantly improving the usability of the retrieved information. Because information access requirements can vary with the type of materials used, we propose to focus our research efforts on a well-defined subset of archived knowledge: the approximately 55,000 papers in the NSF Antarctic literature database and on the collection of users that have an interest in that data.

The selected problem domain has major significance from the perspective of document recognition systems as well. Images recorded on microforms media are frequently low resolution and contain much noise and distortion. Consequently, such images present more difficult problems for preprocessing and recognition algorithms than do images captured from paper documents. The research that we propose will develop techniques that use detailed characterization of algorithms, methods for automatically selecting algorithms with characteristics that suit the image quality, the use of natural language to improve recognition results and prestructure the information to support advanced access methods such as browsing and complex queries, and will make significant contributions to the field of intelligent document understanding. We believe that algorithms to solve all of these problems either already exist or can be created in the time frame of the Digital Libraries program.

We propose to prepare and make available a complete library of information extracted from the microform version of the literature identified above. We will develop models of the way in which Antarctic scientists will interact with this data and techniques to automatically scan, recognize, and organize the information to optimize the utility of the database produced for these scientists. We will make available the conversion techniques to large scale information suppliers and users who could then employ the algorithms developed to organize other similar bodies of knowledge. To make such a large amount of information usable, we will evaluate the effects of recognition errors on text retrieval systems and devise processes based on document and collection knowledge to automatically detect and correct errors. Further, markup can be added to the documents and incorporated into a retrieval system to make the collection more accessible.

We expect that successful completion of this research would make possible the automatic creation of archival databases which would provide users both browsing capabilities and powerful graphic-based queries to retrieve information. Because both the input image domain and the domain of output users is well defined, we will be able to address fundamental research issues such as:

1. models for user interactions with the information,
2. automatic techniques that use these models in the process of extracting and storing information,
3. experimental characterization of document processing algorithms,
4. algorithms that can be optimized to work on specific types of document images, and
5. interaction between search methods and techniques for accessing and representing results of the search.

The research proposed here is grouped according to the technologies that are needed to provide automatic recovery of such information. These areas are pre-processing, character recognition, graphic recognition, and post-processing. There are both commercial and research prototype systems available to perform many of these tasks. It is our intent to evaluate and incorporate as many of the available systems into the proposed testbed as possible. Thus, the final technology incorporated in the testbed produced as the output of the proposed research will be a combination of commercial elements, prototype elements, and new algorithms developed as part of this project.

In order to automate the task of comparing competing algorithmic approaches to each of the recovery steps needed, specially designed experimental environments will be constructed. The design and construction of such test environments has been discussed by [Kanai93].

Experimental "testbed" environments to evaluate both Optical Character Recognition (OCR) and Information Retrieval (IR) systems have been built and have been in operation at ISRI for the past three years (see [Rice93] and [Taghva94a]). These experimental systems are available for the research proposed herein. We expect to build and operate similar systems to compare and evaluate pre-processing and graphic-recognition technologies as part of the proposed project.

Experimental environments to evaluate and compare image processing algorithms, recognition algorithms, and retrieval algorithms require some form of "ground-truth" test data. ISRI has also developed a laboratory for preparing such test data as part of its research for the U.S. Department of Energy (DOE). The facilities of this laboratory will be available to prepare test data for the proposed project.

All development of new algorithms for pre- and post-processing will be done by the research staff at SRI and at ISRI. Comparison testing of competing technologies will be done by the technology assessment group at ISRI. The "ground-truth" test data needed will be prepared by the data preparation laboratory at ISRI. Sample queries and relevance judgment data will be provided by DRI. Traditional bibliographic quality control will be provided by the UNR Library. Agreements with publishers regarding copyright issues will also be obtained and managed by the UNR Library for this project.

In the remainder of this proposal, we first describe (in Section II) the domain that we have selected for study (the NSF Polar Antarctic database) and the rationale for selecting that particular domain. In Section III we present the outline of an automatic system that, upon completion of our research, will be capable of converting and organizing the information in the selected domain. In Section IV we identify the research that must be undertaken to develop technology required for achieving the system vision presented in Section III. Section V describes the rigorous testing environments that we propose constructing to guide and manage the research. Section VI presents our phased plans for distributing the fruits of our research (both in terms of digitized corpora and information capture algorithms) into the relevant user and research communities. We also identify how these communities can provide feedback to influence the course of our own research efforts under the proposed program. The remaining sections cover (in order), the equipment required, project responsibilities, and the proposed schedule and budgets.

II. THE PROBLEM DOMAIN

Since 1951, the Library of Congress, under contract to the National Science Foundation, has accumulated a microform based archival library containing Antarctic research literature. Approximately 55,000 articles have been scanned and catalogued through mid-1993. It is estimated that between 35,000 and 40,000 of these research reports are printed in the English language. The microform library contains a (scanned image) copy of each page of each article in the collection although use of the collection is still subject to copyright restrictions. This Antarctic literature is, of course, only a tiny fraction of the world's scientific literature, but it contains representative titles from all biological and geoscience disciplines. Furthermore, it represents the totality of work which has been conducted on a large and scientifically compelling world region.

An especially interesting subset of the Antarctic research literature can be identified. This subset is the (approximately) 500 articles, pertaining to the so-called Dry Valleys in the McMurdo Sound area. As it happens, the McMurdo Dry Valleys are currently the focus of a Long-Term Ecological Research (LTER) site, being actively studied by the Desert Research Institute, part of University and Community College System of Nevada, also under contract to the NSF. Approximately 25 scientists and students are affiliated with this LTER project, making Nevada home to the majority of practicing researchers in this field. Thus, an ideal subset of documents for study (an experimental/training set) exists and most of the experts on this set are directly available to this project.

Through preliminary discussions with this scientific community, we have identified several modes of access to information which are especially important. First, user friendly systems for efficient browsing and discovery in such a collection are desirable. Also, queries based upon combinations of four basic parameters- "who" authored the study, "what" did he/she measure, "where" were the measurements taken, and/or "when" were the measurements taken- will satisfy the majority of more detailed retrievals.

The following types of automated facilities are thus required:

1. Automated browsing capabilities.
2. Access to information by logical structure.
   (i.e., by name of author or by subject or title)
3. Access to information by geographic location.
   (i.e., information about hydrocarbon concentrations
   on the "west coast of Alaska")
4. Access to information by time period.
   (i.e., information about hydrocarbon concentrations
   on the "west coast of Alaska" before "June 1990.")
5. Collection of information based on above access types
   (i.e., show all papers that used this same methodology in a different
   area of Alaska).

Note that document access by name of author is not supported by most most automated document understanding systems. Access by geographic and temporal based queries is not supported by any current retrieval technology. Similarly, access to sections of documents containing technical (tabular/graphical) information is not supported. We believe that these types of retrieval modes can be developed in the time frame of the Digital Libraries program. The availability of appropriate technology to support the automatic creation of efficiently searchable document databases would significantly help the Antarctic research community and subsequently, be valuable technology for the broader scientific research community.

III. OVERVIEW OF TECHNOLOGY NEEDED

We have seen (in Section II) how the Antarctic research community would like to interact with the information present in the NSF Antarctic research papers database. The key elements of their research strategies can be summarized by a combination of queries of the form: Who (did the research), When (was the data collected), Where (was the data collected), and What (type of data was collected). The objective of such queries is to correlate information in different documents in the collection, and produce a composite view of the knowledge.

Clearly, in designing a system to answer such complex queries, simple string searches of the textual knowledge will not be sufficient. Information present in the hardcopy document must be extracted; separated from unnecessary factors such as page formats and layouts; correlated and connected; and stored in ways that allow easy retrieval and browsing. The volume of documents that have to be processed necessitate the development and use of an automatic end-to-end system. And finally, as the users' experience with the digital library grows, new needs and requirements will be generated. The conversion system must be adaptive and able to grow as requirements grow. Figure 1 shows a diagram of just such a system.

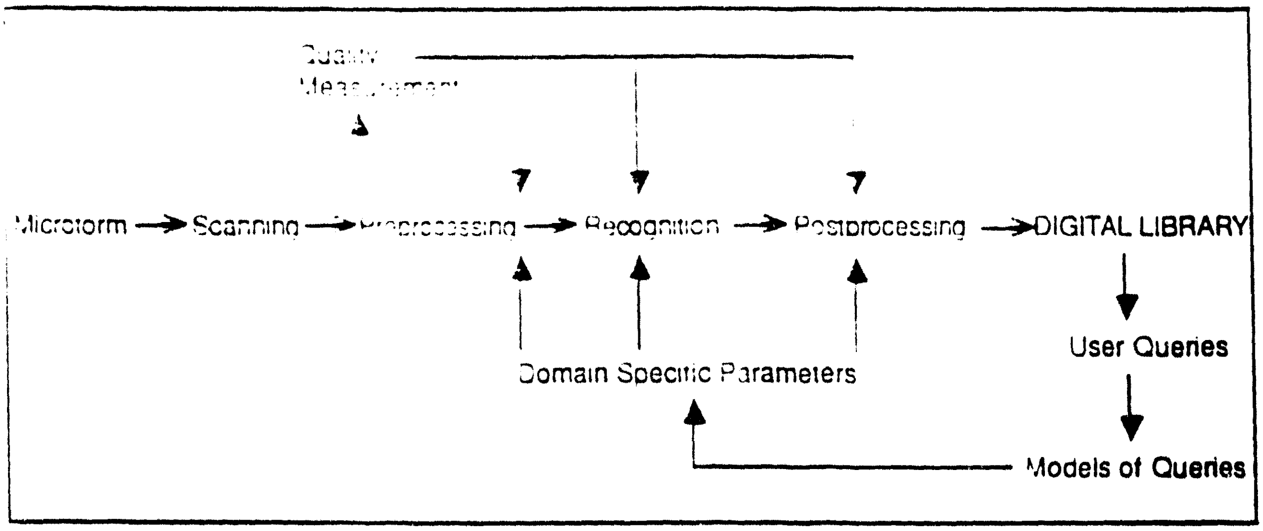


Figure 1. Automated End-to-end Microform Information Extraction System

The general process will consist of scanning the microform documents, recognizing characters and graphics in the documents, followed by postprocessing for error correction and knowledge organization. Users will use the knowledge base (called digital library in Figure 1) in an interactive manner. Their experience with the library will result in improvements in our understanding and representation of the kinds of queries the users make. This in turn will affect the way in which information is extracted, processed, linked, and represented.

Many of these functions are performed by document understanding systems today. In general, given a specific microform collection, current film or fiche handling equipment and current scanning hardware can produce satisfactory digitized images at reasonable cost. The quality of page-images produced is mostly a function of the quality of images on the microform media and not of the scanning equipment utilized [Bradford94]. Thus, in most cases today, automated access to (and use of) large quantities of archival information is limited by current image and text processing capabilities.

Given high quality page-images, current technologies for "Optical Character Recognition" (OCR) produce output character accuracy's greater than 99.5% [Rice93]. The authors of this proposal have conducted a number of recent studies which increase this accuracy to above 99.9% [Rice92], [Taghva93a]. Further, we have demonstrated that, with some additional processing, the residual errors do not significantly affect the performance of current Information Retrieval (IR) systems [Taghva94b]. Thus, at least with good page-image input, currently available technologies will support automated information recovery.

However, significant advances are required in two areas to achieve the vision described by the above model. First, the character accuracy's produced by current OCR systems fall off dramatically as page quality decreases, even though such pages are easily readable by a human reader [Rice92], [Rice93]. Second, current IR technologies do not support several simple types of access to information which are needed.

The next section (Section V) describes our research plan for addressing and solving these critical issues. Section VI describes our experimental methodology that will guide and evaluate results of the research. Section VII describes how the user community and the external research community will interact with our proposed system and provide feedback during its development.

IV. RESEARCH PLAN

Intelligent document understanding is a field that is rapidly maturing. Many commercial systems are readily available that perform various aspects of document understanding with different levels of sophistication. However research is required in several key areas in order to extend the state of the art required to achieve the vision described in Section III.

First, to develop a fully automated document conversion process, optimal image processing algorithms must be selected without any human intervention. For example, at this time, the best way to choose a preprocessing algorithm, such as a binarization algorithm, is by experimenting with the different methods and examining their results [O'Gorman93]. Our proposed research into automatically measuring image quality will lead to the development of schemes for automatically selecting optimal image processing algorithms.

Second, today's document understanding systems provide minimal support for extracting and representing information in the document. We feel that conversion of text to ASCII that can be searched using string matching is just a first step in the process. We need systems that "understand" the information at a level deep enough to assist the end-users find information that they are looking for; combine fragments of information to gain new insights into the information corpus; and organize retrieved information in an understandable manner. For example, our proposed approach finding domain specific phrases using a natural language system and linking them into a hypertext representation will provide users functionality that goes well beyond string searching.

Finally, characterizing the large number of algorithms that are either available commercially or in the research community is a significant area where focused research is lacking. Researchers in communities such as information retrieval, message understanding, and character recognition, understand the value of well organized, accurately ground-truthed corpora. However, many areas of algorithm development are not adequately characterized. An example of such characterization problems can be illustrated by examining the current literature on skew detection algorithms. Almost all of the published algorithms have been tested using small scale experiments. We expect to perform such characterization and place the results into the literature.

This section is broken out into four subsections corresponding to the four processing phases described in Section III. Section IV.1 deals with preprocessing issues such as measurement and correction of distortion in scanned microform images. Section IV.2 describes recognition issues; in particular character recognition for converting and encoding the text portion of the images and issues dealing with recognizing information in graphical elements. Section IV.3 describes postprocessing issues related to the use of lexicons for error correction and automatic markup to support document retrieval. Section IV.3 deals with natural language based systems for extracting and tagging complex information required for structured queries, automatic creation of hypertext links between complex information elements to allow browsing, and the creation of catalogs that show users the results of queries without overloading them with raw data.

In each subsection, we discuss the background (what the problems are in the context of microform conversion), the current state of the art, and the proposed research for solving remaining problems.

IV.1 PREPROCESSING ISSUES

Preprocessing steps serve two roles: they (a) remove scanning artifacts from images prior to recognition, thereby improving recognition rates; and (b) they partition the image into regions based information modality (i.e., text, graphics, halftones) so that each region can be processed by suitable recognition systems. Explicitly or implicitly, preprocessing steps measure characteristics of the image (i.e., image quality or noise) so that the recognition system parameters can be suitably altered.

Typical preprocessing steps consist of binarization, correction of geometric distortion, removal or suppression of photometric noise, and finally page decomposition. In this section (as in the proposed research), we focus only on the latter three areas because we feel that binarization techniques are suitably advanced and within the scope of the proposed research we will not be able to make any reasonable contributions in that area.

IV.1.1 Correction of Geometric Distortion

What is geometric distortion?

Geometric distortions result from improper presentation of the hardcopy to the scanning device. The different types of distortions that can occur depend on the geometric relationship between the microform document and the optics that form the image on the scanner. In scanning devices that move the hardcopy in front of the imaging system, mechanical errors also produce geometric distortions. Detection of distortions typically involves detecting known geometric patterns in the data, analyzing the appearance of these geometric patterns, and computing an inverse transformation that maps the distorted image back to its normal appearance.

What are the sources of distortion?

The most common form of geometric distortion is called skew. It is the result of a misalignment of the document relative to the scanner. In the image, the document axes appear rotated relative to the image axes. There are many commercial and published algorithms for detecting skews as small as a few tenths of a degree (Baird92).

Warping of the original hardcopy relative to the scanner axes can cause higher order distortions. Consider, for example, the copier image of a page near the spine of a hardcover book (Figure). Similar distortions can occur if the microform being scanned does not lie flat in the image plane. Physical damage to the microform (folds or bent corners) can produce similar distortions.

Non-uniform motion of the microform in front of a scanning imaging system can produce geometric distortions that vary with position in the document. Such distortions can give rise to apparent changes in the size of characters in the document, change the aspect ratio of graphics in an unpredictable manner, etc.

What methods can be used to detect distortion?

Typical methods for detecting and correcting distortion are based on recognizing easily detectable geometric artifacts with a known shape. For example, typical skew correction algorithms assume that the baseline of the lines of text in a document are fairly uniform and straight. Skew causes these baselines to rotate relative to the image axis. [Kanungo93] et.al. have published detailed mathematical models of the image formation process in photocopiers and have demonstrated how these models can be used to measure complex distortions.

We can make use of the fact that microforms typically have geometric indexing marks (punch holes) to align the forms with the scanning stage. However, there is no guarantee that the original hardcopy that was used to make the microform was properly aligned. Features (such as the blank spaces between page images in a microform) therefore may not be sufficient for detecting distortions.

What methods can be used to correct distortion?

Most conventional skew correction algorithms rotate and resample the image based on the measured skew. Current image processing hardware can perform complex geometric transformations such as arbitrary warping operations at scan rates. The computer graphics community has developed many algorithms for geometric transformation of images.

What are the research issues?

Given the current state of the technology, the key research issues that we will address are as follows:

1. Mathematical modeling of distortions: As described above, certain classes of distortion models are available in the literature. However, to our knowledge, no systematic effort has been made to model compound distortions. We propose developing detailed physical models of the imaging systems used for scanning the microforms and develop parameterized models of the distortion sources. This will involve making accurate measurements of physical dimensions of the scanning system. Other factors, such as distortions induced by nonhomogeneities of the imaging system (lens distortions and such) will be calibrated by scanning documents printed with known calibration targets such as grids and lines.

2. Determination of measurable geometric features: We will develop a semiautomatic system to identify key geometric elements in the page images. Automatic systems will be used for detecting features such as text lines. Semiautomatic systems will allow the operator to designate repeatable identifiable features such as rules, boxes, and other graphic elements that are typically found in the Antarctic literature. We will develop robust estimation techniques for computing the parameters of the composite distortion model based on measurements from the image. Robust data fitting techniques are being actively studied by the computer vision community (see for example, proceedings of the Robust Computer Vision workshop 1992) and such techniques will be adapted for use in the distortion estimation system.

3. Domain specific distortion correction algorithms: Distortion correction depends on three factors: (a) the modality of the information affected, (b) the processing mechanisms that extract information from the affected areas, and (c) the criticality of the information.

   (a) Text information, line drawn graphics, and photographs will be affected in different ways by the same distortion patterns. Different correction methods will be needed for these three classes of information. For example, preservation of high frequency information will be more important in text regions than in gray level photographs. Preserving linearity of segments in scanned graphs is more important and can be traded off against preserving high frequency components. We will define models for the information content of different types of information regions in a document and experimentally identify the applicability of various distortion removal techniques.

(b) The processing techniques used for extracting information have various degrees of robustness relative to geometric distortions. For example, optical character recognition algorithms are not affected by minor changes in scale because they are designed to operate over a range of type sizes. They are, however, affected by skew or shearing of the characters because shearing can present problems to character segmentation. We will experimentally characterize and then mathematically model the sensitivity of different OCR, line extraction, and other filtering processes, and define appropriate interpolation techniques for each class of operator.

(c) Documents are redundant encodings of information. Text contains stop words that do not contribute to the information content. Rules and other graphical elements are used for stylistic reasons and do not convey information. Word-based analysis can allow recovery of certain errors made during character recognition. However, entries in tables, annotations on a graph, and other similar element have a high information content because they are typically not repeated within the document or present unique relationships with other information elements in the document. We will focus our research on techniques to identify the criticality of the information in various regions of the document and develop methods for concentrating the distortion correction efforts on such regions.

IV.1.2 Noise Removal/Enhancement

Geometric distortions discussed in the previous section are one form of noise that makes a page image appear different from the original document. The second class of noise is photometric noise. There are many sources of photometric noise (or noise for short) that all contribute to the changes in the appearance of the image.

Why noise removal?

Noise removal is required for precisely the same reasons as geometric correction discussed earlier. Noise can cause significant changes in the apparent shape of objects in the image. For example, change in the value of a few pixels from light to dark can change a character "c" to "o". Conversely, a line drawing can change from a series of connected regions to a group of disconnected regions if a few pixels change from black to white.

Further, different OCR devices use different approaches to recognize characters, some devices are immune to a particular kind of noise while others are not. The availability of metrics to measure noise (i.e., speckle and skew) along with document quality (i.e., character brokenness) should make possible experiments to select optimal recognition systems.

Thus the goal of noise estimation and removal is to reduce the noise in the image and attempt to make the image look as close to its ideal representation as possible, and to provide information to downstream recognition processes about optimal recognition algorithms and parameters.

Noise removal methods?

There are three techniques that can be broadly classified as noise removal methods: (a) ad-hoc image processing techniques, (b) techniques based on parameterized noise models, and (c) postprocessing methods that use higher level information. These are discussed below.

(a) Ad-hoc image processing techniques

Most commercial document processing systems perform some measure of noise removal using fixed or adaptive filtering techniques. Common methods are based on median filtering to remove speckle noise, low-pass filtering, and edge enhancement. Recent advances in the field are based on performing noise removal by processing gray level images of the document with more complicated filters such as gray level morphological operators or topological operators [Pavlidis93]. Although these methods may work on the average, they suffer from the fact that many documents contain valid information elements that may resemble noise in other documents. For example, documents where text is printed on a light colored background may present a low contrast or broken text characters that may look like salt and pepper noise; dots in a halftone may have the same size as speckle noise; and other similar confusion s may occur. In such cases, ad-hoc techniques may accidentally delete information that is relevant and should be preserved.

(b) Noise model based techniques.

Such techniques use a parameterized noise model that represents the sources of noise, and uses a fitting process to match the observed image values to estimate the parameters. Once the noise process is modeled, its effect can be reduced, or ideally, eliminated.

In digitizing microforms, noise can be induced in two stages. First, the process of converting the paper document into film can be noisy. Second, the process of converting the film data into digital form can introduce noise.

Noise modeling is an active area of research in the document understanding community today. Two of the most well known noise models are due to Baird (AT&T) [Baird92], and Haralick (UW) [Kanungo93]. Baird's model, which is the most elaborate, focuses on modeling as many of the physical parameters involved in going from an ideal character to its scanned digital representation. It includes terms that model, among other things, the spread of ink on paper, the random jitter of the paper relative to the scanner, and the random thermal noise that is introduced in the scanning process itself.

Work at SRI has focused on a model that captures the variation of the appearance of characters without separating and correlating the variations with specific physical aspects of the scanning process. We have found this method to be more useful than the composition of physical models because the number of parameters required to span the significant changes in appearance are smaller, and consequently, estimating the parameters is possible.

No research has been conducted to empirically or theoretically model the noise that would be produced in the process of digitizing microforms. We propose following a methodology similar to that outlined in the previous section to produce an empirical noise model. We will digitize and scan (a) special calibration patterns, (b) blank and known areas of existing microform documents, and (c) microform versions of the calibration patterns. We will develop suitable parametric noise models based on the digitized values.

Another area of research that we will address is the development of noise models for line graphics and halftones. No such models currently exist. The necessity for such noise models will be driven by the class of queries that the Antarctic research community is likely to make. In general, if the nontext regions are only stored on disk without further processing, no noise removal (hence modeling) is needed. On the other hand, if techniques such as matching images to templates, or matching images to each other is a significant part of the information usage, then noise modeling and removal may be necessary.

[illegible] Postprocessing techniques

In some cases, it may be impossible to distinguish elements of the signal from that of noise. Whether a character is a "c" or an "o" may only be distinguishable based on the word in which it falls. Use of such higher level information, or "context" can play a significant role in noise removal. As discussed earlier, the role of noise removal is to reduce errors in the interpretation of the document information. The role of postprocessing is to remove the errors based on all the information that is available.

SRI has been active in developing techniques for character recognition that are insensitive to noise that is typically found in poor quality document images. Poor quality documents give rise to artifacts commonly called segmentation problems in which individual characters cannot be separated from each other, or in which individual characters break apart into small elements. In such cases, conventional shape-template based OCR systems cannot read the underlying information. We have developed techniques called segmentation-free OCR and segmentation-free word recognition [Chen92, Chen93a, Chen93b] in which lexicons, OCR performance models, and language models are combined to piece together the most likely interpretation of the data. We are investigating extensions of these methods to incorporate natural language understanding techniques to produce further improvements in noise immunity.

IV.1.3 Page Segmentation

Page segmentation is the process that breaks down page images into areas that represent significant information blocks. Typically, a page in the Antarctic database will consist of several significant fields: main document text, graphics, headers, footers, and logical blocks such as author, title, etc.

There are three key elements in page segmentation: separating the page into (usually) nonoverlapping blocks to produce a cover, classifying each block by the type of information it conveys, and finally linking blocks into a logical ordering. Page segmentation is also referred to in the literature as page decomposition and zoning.

Most commercial document understanding systems do a reasonable job of page segmentation when the page structure is simple (i.e., single column of text with some graphics). As the structure becomes more complex (say a Japanese newspaper with different writing styles, read orders, and multiple columns) the problem becomes extremely difficult and no general solutions exist (and indeed, none may be possible).

Issues in page segmentation.

We are interested in the process of segmenting images drawn from a fairly well defined field. The research papers of interest to us in this study are primarily drawn from a handful of journals with well defined styles and structures. Consequently, we will bypass the general page decomposition problem and focus on semiautomatic methods for segmenting pages drawn from a manually defined class of structures.

Text graphics separation.

The first step in page decomposition is the separation of text regions from graphics. This is usually accomplished by image processing techniques that look for the distinctive characteristics of text regions such as horizontal projections, texture [Jain93], or runlength properties [Palumbo90]. SRI has used properties such as edges in a low resolution image [Mulgaonkar90] to achieve similar ends.

The most significant research area that we will explore is the identification of embedded text (i.e., annotations on a graph, vertical place names on a map, floating captions on a photo). We anticipate that researchers using the digital representations of the documents might well want to search for and correlate information presented in such difficult areas.

Classifying nontext regions

Classifying nontext regions as line drawings, bitmapped graphics, halftones, etc. is also a well researched field. We will implement, test, and integrate such techniques in our testbed system.

Read order linking

The page decomposition step described above, segments or zones the image into nonoverlapping regions that are processed independently. Thus, text flows get partitioned into columns, and may, depending on page layout, get fragmented into smaller units.

For searching the text, however, it is important to link the flows together in the order in which they should be read. This process is also referred to as threading. Techniques have been developed for allowing multiple threads to be represented denoting the reading order of elements on a page, for articles that span pages, and to account for subtexts such as boxed items on a page. In general, the problem of determining the reading order has not been solved.

We propose taking advantage of the limited number of editorial and layout styles that will be found in a well defined corpus such as the Antarctic database. We will research four specific techniques to automate the conversion process: rules based on image formats, linguistic rules, directional markers, and analysis of tables of content.

Image Formats: We expect to find a small number of representative image formats in the database of papers to be converted. Most of these will be fairly regular consisting of two columns of text, figures and graphics blocks that are either completely in one column or span both columns, and well defined structures at the start of each article for title, author, and abstract blocks. The flow in such documents is typically linear. We will augment this model by adding other models as appropriate (for example, boxed subtext, running headers and footers, page number locations). We will create a polygonal representation that represents the constant elements on each page at a very low resolution. These low resolution templates will be matched against the overall page image to identify the semantically constant parts of the image. Next, we will indicate typical read sequences within the variable parts of the model.

Linguistic rules: In cases where the read order cannot be precoded correctly, it may be necessary to use character recognition. For any language, we can compute the probability of any word conditioned upon the words that precede it. In the speech literature, this notion is referred to as perplexity of the language. We propose using this notion to evaluate the probability of the first word in each potential column of text given the last few words in a text column under evaluation. The column of text with the highest conditional probability is most likely to be the one that follows the column under evaluation. This metric will likely be very useful when columns are broken within sentences. For columns that break at paragraph boundaries, this measure may not be very reliable.

Directional Markers: In many cases the ambiguity in the read order is most pronounced when articles continue after a gap of several pages. In such situations, physical proximity in the document image is not correlated with read order. However, in such cases, there are typical notations (usually in contrasting fonts) that indicate where the article continues (i.e., 'Continued on page ...', "Continued from page ...". We propose using OCR engines that are keyed to these specific phrases to help in connecting fragmented text into the proper read order.

Analysis of the Table of Contents: The table of contents of any journal can provide a rich source of information about the logical structure of the information in the articles. Pages where articles start, the estimated number of pages in the article, and other information can be extracted from the TOC. This is a rich area of research that has not received much attention in the document understanding community.

IV.2 RECOGNITION ISSUES

Recognition systems take the first steps out of the iconic and into a symbolic domain on the eventual path to a representation of the information. Recognition systems (used in a very general sense) process the cleaned up image in each zone as produced by the preprocessing system. Zone identifications (i.e., text, graphics, halftones) are used to route each zone image to the appropriate set of algorithms. Current document understanding systems typically convert text regions into ASCII using optical character recognition (OCR) techniques, and store graphic areas (such as line drawings, halftones, and gray level photographs) as bitmapped regions. Conversion of line drawings from their bitmap rasters to vector or object oriented descriptions is an active research area (especially in systems for engineering drawing analysis). Identification of significant objects in photographs falls under the broad area of computer vision or image understanding research. There are no commercial systems in which these elements are linked into an integrated process.

In our proposed research efforts, we will focus on two key recognition issues: heterogeneous OCR systems made up of a collection of different OCR engines; and recognition of key semantic elements in graphical data. Clearly, many other issues and system elements have to exist to support the research we propose (for example, image compression techniques) and characterizing and integrating these components will be one part of our activity in producing a testbed system. However, we will focus our research activities primarily on the two areas described above.

IV.2.1 Character Recognition Systems

Perhaps the most important part of any automated information recovery system is the OCR module that converts text images into ASCII code. Errors generated by OCR devices propagate to down stream applications, such as to information retrieval modules or to natural language translation modules. Thus, the performance of the OCR device strongly affects the overall performance of the system. Thus, it is important to investigate methods to reduce OCR errors.

The performance of OCR devices in processing digitized images has been studied by ISRI [Nartker94a]. These results show that one of the most dominant factors in OCR difficulty is the quality of the page image. Image enhancement techniques to improve the quality of the input image are proposed in the previous section. In this section, we propose to investigate error correction techniques which are based on the output text.

Characterization of OCR Devices

The comparative performance of contemporary OCR systems has been thoroughly studied by ISRI staff [Nartker94a]. Important research issues have been identified and discussed by Nartker, [Nartker94b]. At this time, however, no studies have been published which compare the performance of OCR devices in processing microform based images.

We propose to study the behavior of commercial devices, and research prototypes such as the word-based system described in [illegible], using microform based images as input. This study will also collect a-priori knowledge about OCR devices that will be used by other projects in this proposal. (For example, confusion statistics will be used in the smart voting project described in the next section.)

Smart Voting

Bradford and Nartker showed that errors made by OCR devices are highly uncorrelated and that many errors can be corrected by comparing the output from several different devices [Bradford91]. Based on this idea, Rice implemented an automated voting system using several commercial OCR devices. His system generated approximately 50% fewer errors than the best device alone and demonstrated the feasibility of the approach [Rice92]. In this implementation, each device equally contributed to the character error correction process. The results suggest that the number of errors can be further reduced by voting algorithms which use a-prion knowledge about the individual OCR systems. Proposed extensions to be investigated are as follows:

1. Weighted voting based on device performance - votes made by devices that make fewer errors weigh more.

2. Weighted voting based on character confusion's - The conditional probability of a device making a particular confusion, such as 'e' becomes 'c', can be used to weight the vote.

3. Voting with lexicons - select valid words generated by OCR devices.

Segmentation is a major performance bottleneck in many current OCR systems that employ the conventional segment-then-classify approach. SRI is developing a novel approach that performs recognition without first segmenting the text into characters. The approach starts by extracting significant geometric features from the input document image. Each feature then "votes" for the character that could have generated that feature. Thus, even if some of the features are distorted or lost due to character touching or fragmentation, the remaining features can still succesfully identify the character. Preliminary experimental results have shown that this is a very promising approach that can augment and enhance conventional techniques.

We have also extended this research into the recognition of entire words. By describing key words (say in a large lexicon) as a collection of geometric features, the voting process described above can be used to "vote" for entire words. This approch (called keyword spotting) can play a key role in identifying important words in degraded text images where OCR in the conventional sense is not possible.

Global Correction of OCR Errors

A domain specific collection of documents contains words in common. An error correction technique using this knowledge obtained from a set of documents will be investigated. See Section IV.3.1 for detail.

Use of Domain Specific Lexicons

One factor in OCR difficulty is unusual words in the input image [Jenkins94]. As described earlier, many OCR devices rely on lexicons (dictionaries) to make difficult decisions in their character recognition process. Technical terms and proper names are usually not in the lexicon and thus are more difficult recognize. We proposed to study the effects of domain specific lexicons using the dictionaries for Antarctica research compiled by the Desert Research Institute. The key research issues are as follows.

1. Measure character accuracy [illegible] OCR devices with and without domain specific lexicons.

2. Once a string of characters becomes a word in a lexicon, post processing algorithms cannot correct errors in the word. Therefore, it is important to identify the best point in a recognition process to apply domain specific lexicons. We propose to compare the following three points: the character recognition stage, the voting stage, and the global correction stage.

IV.2.2 Graphic Recognition Systems

As described earlier, current document understanding systems process graphical regions primarily as a means to reduce storage requirements. Graphics regions are stored as compressed bitmaps using standard compression algorithms such as CCITT group IV compression for binary data or JPEG compression for image data. Clearly, none of these methods allow end-user queries to deal with graphics (here we use the term graphics to mean any kind of nontextual information) other than noting the presence of graphics and accessing and displaying the graphics on demand.

As we have seen, researchers using the Antarctic database are likely to search the database using four common "query elements": who (did the published work), when (was the work done), where (in the Antarctic region the work was done), and what (was the methodology used). The "who" and "when" questions can typically be answered by reference to information in the text flow. The "where" question is usually answerable by examining the text. However, examination of typical documents in the database show many occurrences of maps that elaborate on or specialize the information in the text. The "what" question is of course the most difficult question and may need analysis of the text and graphics taken together.

Thus, the primary goal of graphics recognition is to produce a representation of the information conveyed by the graphics to enable eventual access to documents and to their graphical components.

Traditional indexing of graphics

In large scale image archives, images are manually coded according to a set of key words that describe the image contents. For example, photographs in commercial stock photo archives would be annotated to indicate that the image consisted, say, of "an airplane landing on a runway at night". Such annotations can be searched using standard text-based queries or can be indexed into a database of keywords and retrieved using Boolean queries.

The limitation of this approach is clearly the necessity to (a) manually encode the descriptors, and (b) estimating the descriptors that end-users would find useful.

Use of Captions

Unlike pictures in a photo archive, graphics in printed documents are not standalone entities. There is a rich source of relevant information available either in the text, or more readily, in the captions associated with the graphic. Identification of captions can simplify the task of cataloging the graphic. In recent innovative research Srihari [Srihari93] has developed a system that couples image understanding of the captions beneath newspaper photographs, and image understanding (face identification) to correlate names with faces. On a simpler level, keyword analysis or deeper semantic analysis of captions could be used to fill in descriptors similar to the traditional manual fields.

Query by Image Content

Neiblack's group at IBM's Almaden Research Center [Ref needed] is developing yet another mechanism for accessing information in images. Called Query by Image Content (QBIC), their technique uses a suite of image features that are extracted from images to describe the distribution of colors, intensities, edges, and shapes present in the image. For access purposes, users can supply crude shape descriptors, either by reference to shape keywords or by sketching outlines of shapes. QBIC executes inexact matches between image descriptors and query descriptors to identify relevant images.

Proposed Research

The approaches described above all have one drawback when it comes to effective use in a digital library system: they do not allow queries that are based on information that is found scattered between text and graphic elements (although the Srihari approach is in the right direction). To illustrate this point consider a hypothetical situation where a particular environmental study is conducted, say on atmospheric ozone. A paper describing the results may focus its descriptive text in describing the methodology. However, the paper may contain a map showing measured distribution of ozone. The map may contain names of cities or places. Later, a user may want to extract all information pertaining to studies in a broad region. The only way to relate the paper to the query would be to understand that some of the place names shown on the map overlap the region of interest in the query.

Our proposed approach to this problem is a model-based one. We propose creating models for various types of geographic information present in graphic images and develop techniques to extract relevant information. We will investigate techniques to drive the model-based recognition using information found in the text. The textual information will be extracted using natural language understanding mechanisms, and will drive the selection and application of appropriate models to the graphics. Once the geographic problem is addressed, we may extend our interest to other graphical elements.

A side effect of this approach is that we will be able to intimately link text and graphics that are related. This leads naturally into hypertext representations that are very important for browsing. Section IV.4 discusses hypertext issues, and in particular, Section IV.4.2 discusses hypertext linking of text and graphical elements. The remainder of this section addresses the research issues particular to the image models that must be created for analyzing graphics and to the issues of structured text such as tables.

Linking text and graphics

In developing the integrated approach, we will initially rely on the structured nature of the documents in the Antarctic database. In archival technical papers, the kinds of graphic images that are typically found are limited to line drawings, graphs and charts, and halftone or graylevel photographs. We will extend the category of graphics to include spatially formatted text such as tables.

Models of Graphical elements

We propose developing models that describe common graphical elements that occur in the documents of interest. For example, maintenance manuals have figures that follow a well-defined style. External views, cutouts, exploded views, and cross sections are constructed from a specific lexicon (or stock graphic elements). Similarly, charts can be thought of as constructed from repeated symbols drawn from a fairly limited lexicon. The team proposes extending research being conducted at SRI on model-based object recognition to incorporate a large collection of known graphics elements. By the recognition of such elements in digitized drawings, the drawings can be converted into an object-oriented description (as opposed to a simple bitmap). NL-based filters will be used to identify key graphic elements expected in drawings by analyzing the text that references the drawing. For example, a figure callout such as "...the pie chart shown in Figure..." can be a cue that the referenced figure consists of wedge-shaped graphics elements. The key research issues that we propose to address are:

1. The identification of models that cover the range of nontext materials in documents of interest

2. The indexing of the models based on external clues

3. Feedback from the graphic elements to tailor the text analysis

4. The analysis of text annotations in the drawings to extract additional information not available from the accompanying text. Specifically we will address identification of place names in map information, identification of generic regions based on collections of place names, and identification of symbology that represents geographic areas of interest such as elevations, flora/fauna, and terrain.

5. The representation and fusion of the information in the two information modalities.

IV.3 POSTPROCESSING AND RETRIEVAL ISSUES

Once the processing stages of OCR and graphic recognition are completed, additional processing is required to render the documents useful for information retrieval applications. Some common information retrieval applications are described below.

1. searching: looking for a particular document, fact, name, etc., which the user has already identified: for example, finding articles related to a specific topic.

2. routing: preprocessing information items for the purpose of sending them to appropriate locations.

3. browsing: navigating without necessarily looking for anything in particular -- such as when reading a newspaper or magazine, or trying to get familiar with a new domain.

4. clustering: grouping together similar items, such as trying to organize large information spaces (e.g., library or mail archives) by topic. This can also be used to determine the degree of similarity between items, such as articles.

which is one means of finding related articles (Botofogo93; Cutting et al. 93; Hearst and Plaunt93; Lewis92; Rose et al. 93].

5. filtering: identifying a priori the desirable or undesirable items, sources, or subjects, resulting in the user receiving selectively disseminated information.

For ease of exposition, we describe issues related to items 1 and 2 in this section. Browsing, clustering, and filtering are described in Section IV.4.

These five functions can be combined for information access and retrieval purposes, but they can also be used for generation and maintenance of hypertext links [Botofogo93]. Corpora used in full-text information retrieval research has traditionally consisted of limited domains or small textual items, but is now expanding to include large or very large collections of text (e.g., see [Cutting et al. 93]). Issues that arise in dealing with very large texts include how to break the text down into manageable pieces (so as not to return a pointer to a whole book in response to a user's query [Salton93a]), and how to derive structure from the text [Fuller et al. 93; Kilpelaainen and Mannila 93; Schauble93]).

Thus, in order to support the desired IR functionality, research is needed in several areas of postprocessing supported by a test system containing a large set of documents. These postprocessing methods range from error correction to information extraction. First, OCR errors can be corrected through the detection and use of global information about the text. Second, the effects of OCR translation errors on retrieval must be accounted for (e.g., errors in acronyms or proper nouns that cannot be corrected by comparison with a dictionary.) Third, the extraction of structure can begin with the annotation of the document's structure using a markup language (e.g., to tag the author, abstract, title, bibliographic references, etc.) Fourth, tools that can perform deeper linguistic analyses can support more precise indexing of documents, improved query analysis, and refinement of a set of documents selected based on a coarse index. Finally, although we have mentioned the creation of hypertext/hypermedia links in conjunction with the recognition of graphics in text, research is also needed to automatically construct hypertext links to other documents, to support database search and retrieval.

IV.3.1 Global Correction of Recognition Errors

Current OCR technologies utilize only local information to translate a single character shape to its ASCII representation. But to recognize document pages commonly found in print, humans typically utilize global information about the page, the document, or the collection. We have developed a set of postprocessing routines which utilize the homogeneous nature of documents and collections, observations about OCR-generated text, and what we know about information retrieval systems, to design a post-processing system to correct errors made by OCR devices [Taghva93a] and [Taghva94c].

The system uses correctly spelled words in the collection and approximate pattern matching techniques to correct misspellings found in the OCR text. The explicit restriction of the lexicon helps alleviate some of the problems of ambiguity found with typical spell checkers. The system uses other global information to determine corrections that are not directly available to an OCR device. It applies document information and a confusion matrix generated from the specific collection to further reduce ambiguity and select the correct character replacement for a misspelled word.

With the large volume of OCR-generated text used in our experiments, we have been able to identify other discernible text qualities. For example, in printed documents, end-of-line hyphenations are quite common. If the document were manually input, the typist would be directed to remove these prior to loading into the retrieval system. Since an OCR device translates exactly what is on the page, end-of-line hyphenations can confuse most retrieval systems. Thus, another function of a post-processing system is to prepare output text for retrieval usability.

We have tuned the postprocessing system to work advantageously with the IR systems used in accuracy experiments described in [Taghva94b] and [Taghva94d]. In each of these experiments, we first compared the results of running a set of queries against raw OCR-generated text to the corrected text. Then, we applied the post-processing system, and in each set of tests, our system improved retrieval.

This post-processing system can be further developed. Currently, it only discovers misspellings which are one edit distance away from a lexicon term. We have found that more errors can be discovered and corrected if we increase this edit distance. Another function that we would like to incorporate is the discovery of inserted or deleted spaces, two of the most common errors made by OCR devices [Rice93]. We also have plans to use additional contextual information for OCR correction. Further, the system can be tuned to work beneficially with the IR system selected for the Antarctic collection.

IV.3.2 Effect of Recognition Errors on Retrieval

An important question when retrieving information from OCR generated text is: "What is the effect of OCR translation errors on retrieval?" We have conducted several experiments designed to evaluate this effect. The results of these experiments have been presented in the following papers [Taghva94b], [Croft94], and [Taghva94d].

The most common use of electronic text documents is in full-text retrieval applications. These applications usually employ an information retrieval system. The papers cited present the effects which OCR text, or in the case of [Croft94], simulated OCR text, have on two different IR models.

In [Taghva94b], two databases were created: a collection of 204 documents (approximately 9300 pages) of OCR-generated text and its corresponding 99.8% corrected version. These collections were loaded into a Boolean logic inverted file-based retrieval system. A set of seventy-one queries pertaining to the document collection were run against each database and the results evaluated. We found there was very little difference between the retrieved results from these two databases.

As we have noted, creating an OCR-generated collection is, in itself, a difficult task. This same kind of experiment would not be practical on collections the IR community considers large. With this in mind, the joint work between the University of Massachusetts at Amherst and ISRI [Croft94] was proposed to determine the level of degradation that would affect larger collections. These experiments degrade a set of standard IR collections with the kinds of errors OCR devices typically make. The statistics applied to each collection reflected the errors that would be generated by OCR devices with average accuracy rates of 96.4% and 87.2%. The results from this simulation support those found in [Taghva94b].

Our most recent experiment evaluating the effects of OCR accuracy on retrieval used an expanded document collection (674 documents), three OCR devices at graded accuracy levels, and the probabilistic IR model, INQUERY, to evaluate the effects of OCR error [Taghva94d]. Again, average recall and precision results showed little difference between the OCR collections and the corrected set. With a probabilistic model though we found certain characteristics of OCR text caused unstable results. In [Taghva94b], we showed that for a simple Boolean system, the problems caused by OCR error could be overcome by redundancy in the document text: but this more comprehensive experiment demonstrated that as the IR model becomes more sophisticated by using word frequencies to construe information, OCR text may not be completely reliable.

We propose to utilize the Antarctic database to conduct additional work on the effects of OCR error on retrieval. The distinctive nature of this document collection can make unforeseen qualities apparent. Furthermore, we have been evaluating the effects of OCR error on, what we call, special recognizers. Special recognizers include such concepts as acronyms, proper nouns, dates, and numerical data. We have recently directed our attention to methods for detecting and correcting these special terms. These considerations are important since special recognizers often appear in database queries. Additionally, they are difficult to correct since they are not usually found in dictionaries.

IV.3.3 Approximate String Matching

Finding all occurrences of a pattern in a document is an essential problem in information retrieval. Computer Science literature is full of clever routines such as the Knuth-Morris-Pratt algorithm for string searching. Since OCR-generated text contains errors, the task of exact match searching must be replaced by approximate matching techniques like [Wu91] or [Landau86]. As part of the proposed project, we propose to investigate with respect to approximate string-matching:

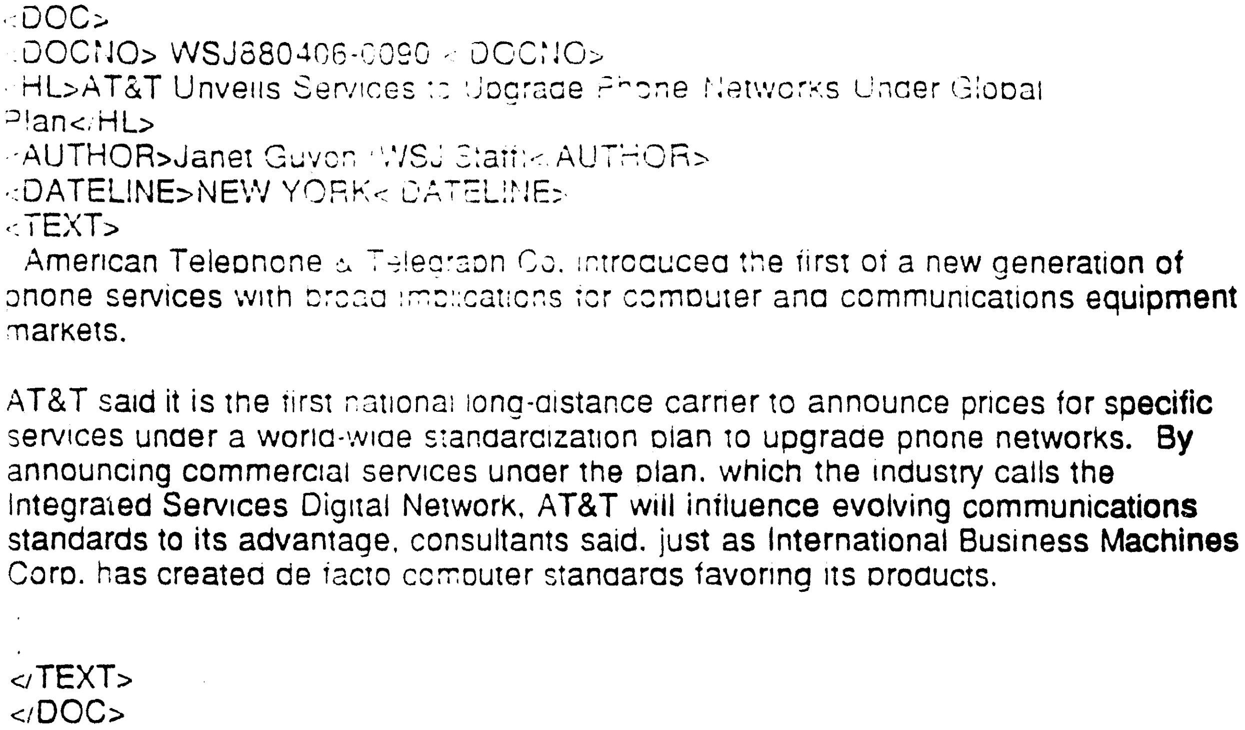
1. Approximate string matching with respect to OCR errors. These errors are naturally different from traditional spelling errors caused by keyboard arrangement. In particular, we propose to develop techniques to identify and correct errors of edit distance greater than 1. [Chen93b] at SRI has developed techniques that characterize OCR output (in terms of a probabilistic confusion matrix) and have developed techniques to measure the "distance" between words in OCR output and words in a lexicon.

2. There are a variety of IR systems with fuzzy matching capabilities. We propose to investigate the performance of these systems with respect to OCR errors.

3. We propose to examine the effects of OCR error on traditional extraction techniques for special recognizers.

IV.3.4 Automatic Markup for Retrieval

Traditional text retrieval systems view a document as a sequence of words and ignore the document's structure. With this view, one can search for documents containing an individual's name, but cannot locate a document written by a particular author, even though this information is available in the document. Similarly, one cannot easily find what an acronym stands for within a document, what papers are referenced by a document, or what words are emphasized in a document even though many papers in the literature [Salton71] point out that exploitation of a document's structure improves recall and precision. Also, most commercial and experimental IR systems require some sort of header information. For example, TREC requires all of its documents to contain SGML-like "tags" so that structured information can be identified by systems involved in these experiments. An example document from the TREC database follows [Harman93].



```
<DOC>
<DOCNO> WSJ880406-0090 </DOCNO>
<HL>AT&T Unveils Services to Upgrade Phone Networks Under Global
Plan</HL>
<AUTHOR>Janet Guyon (WSJ Staff)</AUTHOR>
<DATELINE>NEW YORK</DATELINE>
<TEXT>
 American Telephone & Telegraph Co. introduced the first of a new generation of
phone services with broad implications for computer and communications equipment
markets.

AT&T said it is the first national long-distance carrier to announce prices for specific
services under a world-wide standardization plan to upgrade phone networks.  By
announcing commercial services under the plan, which the industry calls the
Integrated Services Digital Network, AT&T will influence evolving communications
standards to its advantage, consultants said, just as International Business Machines
Corp. has created de facto computer standards favoring its products.

.
.
</TEXT>
</DOC>
```

The items tagged (title, author, etc.) can be considered structured information that is contained in the text but that cannot be identified by the current state of the art IR systems. Generally, these items of information are extracted from the documents and entered into the database manually. Subsequently, this bibliographic information is used by routines like the "SELECT" function of STAIRS [SSS85] in response to structured queries.

Another reason to tag the document is increased document length. Most historical IR experiments have been run against small collections of short documents, i.e., the Cranfield collection [Clevendon62], the NPL collection [Sparck], or the CACM collection [Fox83]. The TREC Conference was designed to evaluate the performance of IR systems against more realistically-sized collections [Harman93]. With this availability of test collections of longer documents, the issue of "passage retrieval" [Salton93b] becomes important. Tagging can mark chapters, sections, and paragraphs. This kind of research would be valuable to projects associated with the digital libraries.

In addition to the bibliographic requirements and passage retrieval, there are also indexing techniques that give higher weights to words based on their position in the document. For example, words found in the title, abstract, or conclusion are given higher weight, since they are more likely to represent the document content. Most IR related research on structured data either presupposes the markup is in place or assumes adding it to the text is an easy task. For reasonable-sized documents from a fairly large collection, this task is prohibitive.

We propose to develop algorithms and routines to automatically mark-up electronically recognized documents. We plan to use the document's physical structure, the document's text, and an SGML document type definition (DTD) to produce a tagged document marking important logical information. These routines will be tested and evaluated on a subset of the Antarctic literature.

The IR community has realized that document representation must go beyond a simple list of non-stop words. Document structure should be preserved and manipulated by the text retrieval system

Macleod90]. The autotagging suggested in this proposal is the key to preservation of document structure. Upon successful markup of a document, the IR system can store, extract, and manipulate the tagged information. So the IR system must be capable of storing not only document text, it must also be able to preserve a document's physical and logical structure. Further, the query language for such a system must have expressive capabilities to handle both document content searching and structure understanding. At the same time, the query language should be English-like with simple syntax and semantics. We believe a natural language interface with a powerful logic query language such as Datalog [Ullman89], will provide a reasonable solution. As a part of this proposal, issues associated with the integration of IR, database management systems and query languages will be investigated.

IV.4 BROWSING AND DISCOVERY ISSUES

To gain the most advantage from a digital DLI archive, users should be able to approach it in several ways. They should be able to extract summaries or articles, query the archive in natural language form, search for and access a particular piece of information, browse the documents in any order, and navigate through the information towards some ill-specified goal. The first two of these capabilities are addressed in the section on FASTUS (Section IV.4.1). The latter two capabilities will be provided by hypermedia links and an information catalog, which are discussed in Sections IV.4.2 and IV.4.3, respectively.

IV.4.1 Natural Language Processing

IV.4.1.1 FASTUS Background

FASTUS is a (slightly permuted) acronym for Finite State Automaton Text Understanding System. It is a system for extracting information from free text in English, and potentially other languages as well, for entry into a database, and potentially for other applications. It works essentially as a set of cascaded, non deterministic finite state automata [Hobbs et al., 92; Appelt et al., 93].

FASTUS has been very effective in evaluations of text-understanding systems. On the MUC-4 evaluation in June 1992, out of sixteen participating systems, only one system performed significantly better than FASTUS, and it had been under development for over five years, in contrast to FASTUS which had been under development for five months.

Moreover, FASTUS is an order of magnitude faster than any other comparable system. In the MUC-4 evaluation [Sundheim92] it was able to process the entire test set of 100 messages, ranging from a third of a page to two pages in length, in 11.8 minutes of CPU time on a Sun SPARC-2 processor. The elapsed real time was 15.9 minutes. In more concrete terms, FASTUS can read 2,375 words per minute. It can analyze one text in an average of 9.6 seconds. This translates into 9,000 texts per day.

This fast run time translates directly into fast development time. FASTUS became operational on May 6, 1992, and we did a run on a set of messages that we had not trained on, obtaining a score of 8 recall and 42 precision. At that point we began to train the system on 1300 development texts, adding patterns and doing periodic runs on the fair test to monitor our progress. This effort culminated three and a half weeks later on June 1 in a score of 44 recall and 57 precision. (Recall is percent of the possible answers the system got correct; Precision is percent of the system's answers that were correct.) Thus, in less than a month, recall went up 36 points and precision 15 points.

The performance of FASTUS in MUC-5 [Sundheim93] on news articles on joint ventures was comparably outstanding. Only two sites achieved better scores (and only one of these was significantly better), and both of these were Tipster contractors who had an 18-month head start. FASTUS outperformed the other two TIPSTER systems, and all nine other non-TIPSTER systems. A Japanese version of FASTUS achieved similar results in the MUC-5 Japanese joint ventures domain.

IV.4.1.2 The Structure of the FASTUS System

The FASTUS system is characterized by five levels of processing. The first four levels operate within the bounds of a single sentence. The last level spans sentences.

Complex Words: This processing step includes the recognition of multiwords, proper names, and possible proper names, locations, dates, times and other basic entities.

Basic Phrases: Sentences are segmented into noun groups, verb groups, and particles. The full complexity of English noun groups and verb groups is recognized. Prepositions, conjunctions, and relative pronouns are also recognized.

Complex Phrases: Complex noun groups and complex verb groups that can be recognized reliably on the basis of domain-independent syntactic information, are identified.

Domain Patterns: The sequence of phrases produced at Level 3 is scanned for patterns of interest, and when they are found, semantic structures are built that encode the information about entities and incidents contained in the pattern. The three criteria that are taken into account in determining whether two structures can be merged are the internal structure of the noun groups, nearness along some metric, and the consistency, or more generally, the compatibility of the two structures.

Merging Incidents: Semantic structures from different parts of the text are merged if they provide information about the same entity or incident.

As FASTUS progresses through these five levels, larger and larger segments of text are analyzed and structured.

This decomposition of the natural-language problem into levels is essential to the approach. Many systems have been built to do pattern matching on strings of words. One of the crucial innovations in our approach has been dividing that process into separate levels for recognizing phrases and recognizing patterns. Phrases can be recognized reliably with purely syntactic information, and they provide precisely the elements that are required for stating the patterns of interest.

IV.4.1.3 Using FASTUS in Document Detection

It is impractical to respond to an ad hoc query by applying the techniques of information extraction to every document in a very large collection. However, there are three areas where information extraction technology can support detection: (a) indexing documents before they are archived, (b) query construction, and (c) refining the precision of the top n documents.

Each of these is discussed in turn in the next three sections.

(a) INDEXING DOCUMENTS

When documents are received, they are indexed. This is an appropriate point to apply natural language analysis techniques to them to determine the principal patterns of interest they exhibit.

This is most obviously [illegible] for known profiles defined by end-user researchers using the collection of data. Assume these have already been expressed as abstract template types and associated sets of patterns have been constructed. All such patterns would be run against the text, and for every match, the appropriate users would be notified. But linguistic analysis at indexing time should also improve the efficiency and precision of ad hoc queries, by allowing more fine-grained indices.

Natural language analysis in indexing allows a range of patterns to be identified and to serve as indices to the documents. FASTUS can aid in indexing at the phrase level, at the level of domain patterns, and at the discourse level of merging entities.

The phrase-level analysis of FASTUS enables the recognition and normalization of the names of organizations, persons, products, locations, and so on, so that documents can be indexed on these. More sophisticated techniques are possible as well. It may be that head nouns of noun phrases are better indicators of relevance than words in general. It may be that subjects of sentences are better indicators. Previous experiments have shown that using sequences of nouns as keys improves performance. It may be that the use of full noun groups, that is, the noun phrase up through the head noun, improves performance even more. These hypotheses suggest experiments that can be conducted using only the phrasal analysis of FASTUS, requiring no knowledge of the domain.

Although all the domain patterns required by ad hoc queries cannot be anticipated, it is likely that there are broad areas of commonality that can be exploited. This possibility will be investigated by examining the 150 available TREC topic descriptions to determine the generalizations they exhibit, encoding them into finite-state patterns and using them in document detection. Patterns that are found to be especially effective can then be searched for whenever a text is received and indexed.

The importance of a concept in a document is generally related to the frequency with which the concept is referenced. However, the references can become quite spare with the use of anaphoric devices. An article about sanctions against South Africa might contain the whole phrase just once, and subsequently refer to them by "sanctions" or "the measures" or even "them". To the extent that we can resolve these references using the techniques for merging entities in FASTUS, we will have a much better measure of the centrality of a concept in a document. Concepts that are referenced more will be given greater weight as indices. This technique can be used both in indexing and in precision refinement.

To summarize, we will perform the following experiments:

We will determine whether detection is improved by searching on full noun groups, weighting head nouns more heavily, and weighting subjects more heavily.

We will determine to what extent a hierarchy of common, general patterns is exhibited in TREC and TREC-like topic descriptions and whether indexing on these improves detection.

We will determine whether weighting concepts by the number of references a document makes to them, having resolved coreference, improves detection.

(b) QUERY CONSTRUCTION

There is much that linguistic analysis can do to improve queries. Currently, detection methods use only co-occurrence information about words, but these co-occurrences can often be accidental. In fact, accidental co-occurrences may be the source of many of the false positives in detection.

Linguistic analysis of a topic description can determine when the co-occurrences represent predicate-argument relations, and can thus provide a more focused query to search on. In particular, linguistic analysis can determine when relationships occur within the scope of negation or other modal or epistemic operators in topic descriptions, so that the query will be right on the concepts that are searched for. Where the predicate-argument relations are common enough that documents have been indexed on them, they can be part of the query that is used in the initial retrieval from the full document collection. Otherwise, they can be used in reprioritizing the retrieved documents.

Linguistic analysis may also be used to distinguish between the significant concepts in a query and the concepts that are not very indicative of relevance. For example, in the TREC topic

> \>>A relevant document will describe countries supplying Japan
> with agricultural products and will specify the quantities.<<

we should not search on the words "relevant" and "document". A more interesting problem is presented by the word "specify". The least we will want to do is to downgrade the weight on this term or eliminate it entirely. We might refer to words like "countries", "supplying", and "Japan" as content words and words like "describe" and "specify" as characterization words. The latter characterize relevant documents rather than indicate their content. In our research, we will examine the structure of the TREC topic descriptions to determine how content and characterization words most commonly occur.

Characterization words like "specify" and "identify" present a more difficult problem, however. How can we determine automatically whether a document specifies or identifies an entity? What counts as an identification or a specification? In general, this is a very hard problem and we cannot hope to solve it within the scope of this project. However, it is likely that there are many types of entities for which we can define what counts as a specification. Quantities are one of them. "10,000 metric tons" is a specification while "a lot" is not. Where entities have names, as with countries and companies, the names count as identifications. However, if relevant documents must "specify the means of debt restructuring", it is not obvious what would count as a specification. In our research we will attempt to identify as many cases as possible where a precise definition of "specification" can be devised.

(c) PRECISION REFINEMENT

Linguistic analysis cannot be applied to every document in the collection every time a new query is formulated. The FASTUS system provides very fast natural language processing, but not that fast. But if the detection component first narrows the document set down to a thousand or so ranked documents that are highly likely to be relevant, then FASTUS can be applied to each of these documents, in rank order, and the set can be reprioritized on the basis of this deeper analysis.

An experiment performed in connection with MUC-4 indicates that this could be very effective [Lewis92]. The texts in the MUC-4 corpus were found by doing a keyword search in the much larger FBIS corpus, and the result for one test set (TST4) was a set of documents, of which about 55 were relevant. The MUC-4 systems were considered merely as detection systems, not as extraction systems, and a document was considered relevant if the system found an incident. The performance of FASTUS on this test was typical of the most effective MUC-4 systems. It achieved a recall on TST4 of 93 and a precision of 82 , a substantial improvement over keyword search.

Reprioritizing can be done using FASTUS. If domain-dependent patterns have already been developed for this topic, they can be run against the retrieved documents, and reprioritization would be a function of the matches found. Even if patterns have not yet been developed, precision refinement can play a role. The user can scan the highest-ranking documents, in which the key words used in retrieval have been highlighted. Perhaps the user extracts information for a template, or perhaps he only rates the documents as relevant or irrelevant. In either case, he marks the phrases in the text that indicate relevance or irrelevance. The sentence that contains the phrase is analyzed by FASTUS to discern a pattern. As patterns are induced, they are applied to the rest of the retrieved documents for reprioritization, giving us a kind of fine-grained relevance feedback mechanism.

To conduct this investigation, we will choose a number of queries for which roughly equal numbers of relevant and irrelevant documents are among the high-ranking documents returned by a keyword detection system. In a later stage of the research we will also look at queries for which the detection component gives very poor results, to determine how the system could be improved on these queries. We will examine the high-ranking relevant and irrelevant documents and the low-ranking relevant documents to determine what patterns could have resulted in a more nearly correct ranking. We will determine what generalizations over these patterns are appropriate, and either how they could be constructed automatically at query time or how they could have been constructed conveniently by the user, perhaps using a library of patterns.

A preliminary examination of this data suggests that the primary reason for high-ranking irrelevant documents is that relevant events or relationships are embedded in modal or epistemic operators that make them irrelevant. For example, if the topic is government retaliation against terrorists, then among the sentences that do but should not trigger retrieval are the following:

> The U.S. will not retaliate against terrorists.
> The U.S. should retaliate against terrorists.
> The U.S. must decide whether it will retaliate against terrorists.
> The U.S. has plans for retaliating against terrorists.
> The U.S. is discussing how to retaliate against terrorists.

If this generalization holds up on further investigation, then retrieval may be made much more precise by searching for these contexts around the key concepts and filtering these matches out.

IV.4.2 Hypertext

The following list shows the desirable characteristics of an information database that make it a good candidate to be represented in hypertext format. We believe the polar Antarctic database and its intended use fits this list.

- The textual/multimedia information already exists, either in computerized or computerizable format. We have already seen that the complete knowledge base exists as microform documents and is available for conversion.

- Users will benefit from non-linear access to data. As described in Section II, researcher would like to browse through the data, pulling together information from disparate documents in the collection to develop a composite picture of the knowledge that is encoded in the corpus.

- Without restructuring, the present information system is computationally unwieldy in size and/or complexity. The corpus consists of approximately 40,000 documents, with each averaging 25 pages. This is an unwieldy structure for users to navigate.

For the majority of hypertext systems that have been developed, the creation and maintenance of links have been performed manually; this is a tedious, labor-intensive, and error-prone procedure. Because of this characteristic and a desire on the part of hypertext proponents to extend the node-link model to include various user types and purposes, current research is making significant progress in the automation of link functions. In this respect, hypertext has benefited from recent developments in information retrieval and artificial intelligence techniques (references). Such technologies have resulted in more accurate, complete, and faster creation of links, and in more intelligent and subtler interpretation of relationships between nodes.

We will research the issues involved in developing hypertext links automatically from OCR-generated databases by developing a research prototype of a system that supports the development and testing of different link creation algorithms. We expect that users will have to be involved in verifying the link creation, so the prototype will involve user interface development. The research prototype will be developed in four steps: performing requirements analysis and knowledge acquisition; deriving the taxonomy; selecting and/or adapting an existing algorithm for automatic creation of links; and developing a prototypical GUI for use with the system.

Because of the high overhead in creating manual hypertext links, current hypertext efforts today are limited to corpora such as encyclopedias that are updated relatively infrequently. Since our research is directed towards the automatic creation of hypertext links, we will have the ability to augment the collection by adding new documents that will be linked into the existing hypertext information.

For all of the application-oriented issues described (e.g., query formulation, hypertext usage, and database catalogs), we will be best able to serve the needs of the users by determining their needs in accessing the polar Antarctic database. We can determine, for example, the following: (1) what functions need to be supported (e.g., full text retrieval, creation of personalized hypermedia links; previewing of image data, etc.); (2) what, if any, performance parameters the users feel should be met; (3) what types of links would be most useful, how they should be depicted, what actions should occur when a user traverses a link; and (4) degree of granularity to be supported by links and catalog; for example, should the base unit be an article, paragraph, sentence, word, or something completely context-dependent?

Once we have determined what needs the users have, we can develop a taxonomy of link types to support the link creation. We will analyze the types of postprocessing discussed previously (e.g., tagging of the document, natural language analysis, graphic recognition) to determine what information they can provide in link creation, then review available algorithms for link creation to determine their applicability. (Most of the current algorithms are in the research stage.) We will study several algorithms and acquire one or more, which we will adapt to our domain, corpus, and application.

Finally, to facilitate testing of the created links with users, we will provide a prototypical graphical user interface (GUI) for navigating the document database. To the extent possible, the GUI developed here will mesh with that developed for the data catalog (see next section). At this time, we cannot theorize too much about the details of the GUI, since certain characteristics of it will necessarily depend on such factors as the target platform, available GUI tools for the platform, results of the requirements analysis, etc. However, there are certain well-defined cycles in the normal course of development of any GUI. These include:

- developing a storyboard to understand the process by which users will accomplish their goals
- showing the storyboard to prospective users to get their feedback
- revising the storyboard per this feedback

- implementing the GUI design in software on the target platform

- conducting usability tests in which representative users perform representative tasks

- revising the on-line GUI as appropriate

In the polar Antarctic database, it will be important to link together the text and the graphics. For example, figure callouts in the text have to be linked to the figures themselves to allow the storage system to efficiently organize the information and permit the user to easily access related information that is described in both text and graphics form. The only significant effort in linking text and graphics that has been reported in the literature has been by Srihari [Srihari93] in which captions in newspaper photographs were interpreted to identify the nature of the photograph.

We propose to study the issue of automatically linking text and graphics fields in a mixed document to create a hypertext representation. We propose to develop filters that can rapidly identify key concepts in the text to enable linking, and to develop techniques for linking information extracted from the text with corresponding information available in graphic images in the same document.

We propose using, extending and applying SRI's FASTUS technology as filters to extract information that describes or otherwise relates to the graphic elements in the document. Typical candidates are text such as picture captions; phrases such as "...as shown in Figure..."; and spatial and relational words such as "...on the top left corner of Figure... ." The key research issues that we propose to address are as follows.

1. The identification of common phrases and structures that are used to describe pictures and graphics in any given corpus

2. The effect of partial recovery of key phrases on the usefulness of the resulting links

3. The ability to construct filters for text in a new domain in an automatic or semiautomatic manner.

IV.4.3 Building an Information Catalog

Scientists doing research will want to create their own set of relevant articles from the complete archive. Our objective in this task is to perform research to support that goal by examining search and retrieval operations on the database (containing, at this point, text and scanned images, graphs, etc.) that will permit individual users to customize it to their own uses. Our main focus is to examine the utility of creating and maintaining a local cache for retrieved data that can be used as an on-line reference base for future accesses.

First, we will build an object model to support storage and retrieval of the source documents, using a standard approach to object modeling (such as described in [Rumbaugh et al. 1991; Shlaer and Mellor 1988]). Then, the preprocessed documents will be compiled into an object-oriented database (OODB). In an OODB, the database objects have various properties including classes, class hierarchies, inheritance, encapsulation, object id, and interobject links [Cardenas and McLeod90; Catell91]. To manage the database, we will purchase an off-the-shelf database management system (DBMS), such as Versant's Object Database, Objectivity's O/DB, ITASCA's object database, or Servio's GEMSTONE. The DBMS will need to support persistent storage, concurrence control, short and long transactions, location transparency, and object migration, among other features [Joseph et al. 90; Zdonik and Maier 90]. We will use the DBMS to define the document objects, structures and organization.

In the next stage, we will devise an index (similar to a database dictionary) to indicate information about each file in the database, such as identification, source, format, media type, length, size, create/modify dates, and related keywords. The index will serve as a reference point for a browser, which researchers can use to locate information of interest to them. It will provide an interface for the users, and will allow them to search, query, access, and retrieve the contents of the polar Antarctic archive. The model of the database will reflect the database schema, with respect to how the items in the database relate to one another (e.g., hierarchically). The results, or "hits," of a search can be presented to the user as iconic objects through a graphical user interface developed for the purpose of this application.

We will use the commercial database's fourth generation language (4GL: e.g., SQL), host language interface, or visual toolkit to develop query semantics that will allow searching and browsing based upon object attributes, relations, or types. We will also implement a set of simple capabilities that will allow users to sample large data types.

A user will access the database and perform a search by issuing a natural language or 4GL query. The query is translated and ported to the document DBMS, where they are matched against the index for each entry. The name, id, and other select information for each matching article is displayed as fields of an entry in the "catalog" on the user's desktop. Each image, chart, or map associated with each article is displayed as an icon for that article's entry. Icons are shown in their relation to one another: items that are sequential, for example, or constitute subsets of a superset, will be so indicated: items that might not exactly match the search parameters but are closely related to those that do match can also appear.

Users review the results of their searches, and if they would like to get a better look at something, they can choose to "sample" it. Sampling an image entry involves showing bits of it in less than complete form. Sampling will be media-dependent: that is, if video data were available, sampling a video could involve showing every tenth frame, or the first 60 frames: sampling a text file might mean sending over such information as an abstract, introduction, or outline (if available). The advantage of this process is that it reduces the load on the network, and the reviewer can get an idea of whether or not the item is worth copying over to his workstation's local memory for his use.

Users can save the catalog, which contains the results of one query, to a file for later use. They can also modify the original query, resulting in the issuance of a new catalog, which they can compare on a side-by-side basis with the first catalog.

If the user decides that an item in the catalog is of further interest to him/her, he/ she can copy it over to a local disk. Users have the option of copying over the entire document file, or specifying which parts they want to copy over, such as the first 10 pages, or first 10 Kb; alternatively, they could choose to retain the reference item without actually copying over any data. This selection process, known as "culling," minimizes the amount of disk space used. In addition, since the original query, query results, and source locations (and, optionally, samplings or full files) are saved in the users' files, subsequent work or further refinements of the query can be performed without having to go through the search and sampling process. Once the data has been retrieved to a user's local disk, it can be edited using whatever local packages are available: for instance, having viewed the entire file, someone may want to include only certain segments of the file in the interests of space or relevance.

This system will provide the following advantages:

1. Users can cull through large amounts of information and select a limited subset for their use.

2. Catalogs can be saved for future use, preventing the user from having to reissue queries and repeat the search process.

3. Both network bandwidth utilization and local storage requirements are minimized.

V. EXPERIMENTAL METHODOLOGY

In order to select an acceptable set of image and text processing routines to provide the end-to-end capture and delivery technology needed, we will require several sets of test data and a system to automate experiments, using this data as input, to compare competing algorithms and ideas. Thus, our methodology is concerned with the selection and preparation of test data and with the design of an experimental environment which automates all comparison testing activities.

An automated system is needed to make possible experiments with large, statistically significant, quantities of data, to insure that such experiments are reproducible and to eliminate human error and bias. The necessary components of experimental systems to automatically compute measures of performance of document processing technologies have been presented by the authors in [Kanai93]. Briefly, the following components must be provided:

1. a collection of appropriate test data (images or documents),
2. for each test image or document, a "ground-truth" representation corresponding to the property to be measured,
3. algorithms for comparing the processed output (for each test system) with the corresponding "ground-truth" data, and
4. algorithms (or metrics) for analyzing deviations between the output and the corresponding truth data.

Over the last three years, ISRI has built and operated systems to measure the performance characteristics of both OCR devices and Information Retrieval systems [ISRI93]. These systems are available for the research proposed herein.

V.1 Preparation of Test Data

For preprocessing and recognition technologies, all test data are images of document pages. We plan to obtain and to scan original hardcopies of several test documents. We wish to compare the quality of images from hardcopy pages with those on microform in the NSF library.

In order to test competing preprocessing and recognition algorithms, we plan to select a uniform random sample of 1,000 pages from the approximately 1,000,000 pages of NSF English language documents. (40,000 English documents x 25 pages per document) If the average page were to contain 2000 characters, each experiment with this dataset would involve 2,000,000 characters. It is important to select a set with a representative mix of document features (figures, maps, tables, as well as mainbody text) against which recovery is desired.

The selection, scanning, and zoning of images as well as preparation of the corresponding "ground-truth" data will be conducted by the ISRI data preparation laboratory. This lab currently produces about 200 pages of "ground-truth" data per month. We expect this 1000 page sample can be prepared in about 8 months.

In order to test algorithms to structure document databases and algorithms for browsing and retrieval, an appropriate selection of entire documents will be required. As mentioned in Section II, a very appropriate document set exists. The 500 articles pertaining to the Dry Valleys in the McMurdo Sound area of the Antarctic will be scanned and recognized. As with the page-based data, the ISRI data preparation laboratory will prepare these documents.

A main component of the information retrieval test system developed at ISRI over the last three years is a set of test documents. These documents are part of a large DOE document database which has been described by [Nartker92]. The collection consists of approximately 2,600 documents (104,000 pages) together with their corrected ASCII text and original page images.

We have also obtained from the DOE a set of 120 example queries which are appropriate for this document database. To provide the "ground-truth" for this test dataset, we have prepared the corresponding set of relevancy judgments for each query. This database is also available to compare competing retrieval algorithms. Of course, the 500 articles pertaining to the McMurdo Dry Valleys will provide the most applicable tests.

V.2 Measures of Performance

For preprocessing and recognition technologies, there are a number of possible measures of performance. First, each competing algorithm can be evaluated by measuring the recognition accuracy produced. Thus, one class of metric for preprocessing algorithms produces "character" accuracy, "word" accuracy, etc. Of course, preprocessing algorithms which uniformly improve character accuracy output from an OCR device are highly desirable.

ISRI has also developed several new measures of OCR performance [ISRI93]. Our current test system will measure non-stopword accuracy, phrase accuracy, marked character efficiency, and cost of automatic zoning.

We plan to experiment with other performance measures as well. One measure we have not yet considered is throughput. We have considered that device throughput is much less important than device accuracy. Other types of metrics might include logical decomposition or noise or page quality.

Over the past three years, ISRI has used standard measures such as precision and recall to evaluate IR systems in the presence of OCR errors. These measures will be used to compare competing algorithms in pre- and post processing.

V.3 Components of the Test System

The actual preprocessing/OCR test system is a software system which accommodates the installation of competing devices and controls their operation automatically during each test. A main feature of this program is a vendor independent interface which normalizes the output from each competing algorithm to make the direct comparison of output possible. The test system is implemented as a set of Unix shell commands [Rice93].

Because ISRI conducts annual OCR Technology Assessment tests for the Department of Energy, it obtains the best OCR technology available from each competing commercial vendor each year. These systems are available for the proposed research. Unless some new prototype system is introduced which provides an order of magnitude improvement in performance over current commercial systems overnight, ISRI is in a position of selecting the best available OCR device for any given set of data very quickly. Choosing an optimal combination of technologies to provide the end-to-end abilities needed will require more time.

To compare information retrieval systems, we have installed one Boolean logic positional inverted file system (Basis Plus), one probabilistic retrieval system (INQUERY), and one vector based system (SMART) in order to conduct model based retrieval experiments. ISRI plans to acquire several other retrieval engines (based on different models) within the next few months. These will be available for use in the proposed project.

VI. TESTBED FACILITY

In Section V.1, we described two major sets of test data sampled from the NSF Polar Antarctic literature archieve. The first set will contain 1000 page-images randomly selected from the archieve and the corresponding ASCII text for each page. The second set will contain 500 documents associated with the Dry Valleys in the McMurdo Sound area of the Antarctic, a set of queries pertaining to these documents, and relevance judgements associated with each query. Both datasets will be available by the end of the second year of the project.

We propose to make these datasets available to character recognition and information retrieval researchers. (In the case of the image dataset, 500 pages will be made public during the first year, as a "training" set. The second 500 pages will not be distributed in order to make possible independent evaluation of all new recognition systems produced.) Each year, beginning with the second year, as part of the ISRI technology assessment test activity, we will publish a comparison of the performance of all recognition systems submitted using the 500 page "test" dataset. In addition, ISRI will devote a special session at the "Symposium on Document Analysis and Information Retrieval," (SDAIR) held each year in Las Vegas, to papers which utilize these datasets for research .

At the end of the project, the Digital Library of Polar Antarctic research literature produced as the final product of this research will be installed, maintaned and managed on a file server at the Desert Research Institute in Reno, Nevada, or another contractor site designated by NSF. Because the DRI facilities in Reno are accessable via a T1 line to all Internet sites, researchers anywhere with Internet access will be able to utilize the system directly. In addition, dial up access will be supported for those with access to a telephone but not to the Internet. For researchers with neither Internet nor telephone access, we plan to make available a CD-Rom version of the system.

UNLV and SRI will equally share the rights to license any software developed on this project, except software which is a FASTUS derivative. FASTUS is a software system developed in Lisp by SRI International using SRI IR&D funds (i.e., developed exclusively at private expense per FAR 252.227-7013 (a)(12)). Rights to FASTUS and FASTUS derivatives will be retained by SRI. If FASTUS is needed by the Government for use with this program, it will be delivered under Restricted Rights subject to SRI's standard license agreement. In addition, upon request, we will deliver any software developed under the proposed project to the Government with rights to use the software for Government purposes only and to distribute the software to other participants in this program for their use only in this Government program, subject to a license agreement.

VII. EQUIPMENT REQUIRED

With the exception of microform scanning equipment, all equipment necessary during the first year to digitize and prepare test data is already available in ISRI laboratories in Las Vegas, Nevada. Equipment needs for the later years of the project are difficult to predict.

For the first year, funds are requested to acquire microform scanning equipment, additional disk storage equipment, a prototype file server, two workstations for Antarctic researchers, and four workstations for graduate students (see Section X). An appropriate network of Unix workstations and file servers currently exists and is available for the research staff at both ISRI and SRI.

Funds are requested to acquire several competing commercial software systems, including both image preprocessing, text processing and information retrieval systems (see Section X). All competing OCR systems are already installed for experimental research at ISRI in Las Vegas.

VIII. PROJECT RESPONSIBILITIES

All development of new algorithms for pre- and post-processing will be done by the research staff at SRI and at ISRI. Comparison testing of competing technologies will be done by the technology assessment group at ISRI. The "ground-truth" test data needed will be prepared by the data preparation laboratory at ISRI. Sample queries and relevance judgment data will be provided by the DRI. Beginning in year two, continuous subjective evaluation of the prototype Digital Library system will be provided by LTER scientists at DRI. Traditional bibliographic quality control will be provided by the UNR Library. Agreements with publishers regarding copyright issues will also be obtained and managed by the UNR Library for this project.

Through DRI, LTER staff will contribute their expertise to the proposed Digital Libraries project in three areas: 1) building the prototype document database, from hardcopy materials in researchers' personal libraries; 2) defining the "ground truth" for this prototype database via query/response scenarios against it; and 3) testing the proposed document retrieval system, in a real-world setting, as it progressively develops.

UNR Libraries staff will also contribute their expertise to the proposed Digital Libraries project in 4 areas: 1) documenting the current library utilization and research methods of the DRI LTER scientists in a "base-line" study; 2) providing independent quality control of the prototype database and query/response scenarios; 3) assisting with human factors (interface design, performance measurements, etc.) of the proposed retrieval system, as it progressively develops; and 4) assessing the sociologic accomplishments of the project, from a library perspective, in a retrospective study compared against the base-line.

IX. EXPECTED ACCOMPLISHMENTS

In the previous sections we have described the project goal, which is to develop techniques for automatically converting a corpus of microform documents into a usable collection of information. We have described the corpus of interest and the user community in Section II, outlined the expected system in Section III, and defined our research agenda and experimental methodology in

Sections IV and V. In Section VI we showed how we would organize and open up our testbeds to other researchers.

This section describes significant milestones that (a) will define our progress, and (b) provide a mechanism for monitoring the significance of our accomplishments. In Section IX.1 we identify the target that we will be aiming for at the end of the four years of NSF funding. Section IX.2 shows the intermediate (annual) milestones that we propose establishing to keep the project on track.

IX.1 Four Year Goals

At the end of the four years of funding, we will have performed significant research as well has have a complete testbed with a digitized corpus of approximatly 40,000 documents and an active user community performing research using the digitized corpus. We will measure our success by the number of contributions made to the research literature. We expect to have achieved the following:

* In the Information Retrieval Field

We expect that over the course of the project, we will publish new algorithms for automatically postprocessing information from scanned documents to (a) globally correct recognition errors, (b) create automatic markup for fields required for document access, (c) natural language based document indexing, (d) querying, (e) precision refinement, (f) hypertext link creation, and (g) information cataloging.

We expect to publish results of experiments in information retrieval characterizing the effect of OCR errors on precision and recall, as well as evaluating the effect of OCR errors on natural language understanding. These results will be useful for researchers studying new algorithm directions and engineers constructing special purpose systems.

* In the Document Image Understanding Field

The proposed research in image preprocessing and recognition will result in new algorithms for (a) geometric distortion correction, (b) modelling and noise suppression algorithms specifically directed to microform images, (c) trainable systems for page decomposition, (d) trainable, smart voting schemes for combining the performance of disparate OCR engines, and (e) techniques for extracting key information from graphics.

We propose publishing the detailed characterization of preprocessing algorithms including character recognition error rates, page decomposition failure modes, and noise models in an effort to stimulate further research in the document understanding community.

* In the User Community

One very important measure of the value of our proposed research will be the number of references in new Polar Antarctic research publications that reference the digital collection. We expect that as researchers in the community are exposed to the new research capablities provided by the online collection, their use of this resource will grow. A measure of the growth will be the number of citations that directly reference the resource. Also, as the value of the resource becomes recognized, the number of researchers that use the collection will also grow, providing another quantifiable measure of value.

* In the Information Providers Community

We expect that by the end of the four year project, we expect to have entered into direct negotiations with information providers who would view the automating information conversion system as a resource that they could use to rapidly bring other information compendiums to market.

IX.2 Annual Milestones

We have identified annual milestones that will chart our progress in the four year program. These milestones are broken out into two areas: (a) research program, and (b) prototype functionality.

IX.2.1 Research program

During the first year of the research, we will focus on the following elements:

a. Preprocessing issues: We will study distortion models (both geometric and photometric) and techniques for estimating these models from digitized documents
b. Recognition issues: We will study smart voting schemes, extending current research ongoing at UNLV.
c. Modelling of User Needs: We will interview users and evaluate their needs and research methodologies. producing storyboards or other mechanisms to capture the knowledge.
d. Information Retrieval Issues: We will study the nature of markup tags required based on the user needs, and mechanisms for automatically acquiring such tags by postprocessing the documents.
e. Algorithm Characterization: We will perform detailed characterization of algorithms in all categories based on microform images. and identify areas where current algorithms are weak.

During the second year of the research. we will build on the accomplishments of the first year, expanding our research to:

a. Preprocessing Issues: We will augment our document understanding capabilities with development of trainable page decomposition and zoning systems.
b. Recognition Issues: We will study graphics processing techniques and identify information that can be extracted for indexing and markup purposes. We will integrate word recognition and domain specific lexicons into the recognition engines.
c. Postprocessing Issues: We will perform and publish results on characterization of the interaction between recognition error rates and information retrieval precision and recall.
d. Browsing Issues: We will start examining natural language based techniques for extracting and linking key information in documents.
e. Results Sharing: We will host the first of three annual workshops geared to discussion and comparative evaluation of research results based on the digitized polar antarctic database documents.

During the third year we will focus on:

a. Postprocessing Issues: We will evaluate robustness of automatic techniques developed in the previous years and integrate feedback from the preliminary users of the conversion system and converted data to improve the relevance of the system to the end-user community.

b. Browsing Issues: We will develop techniques that integrate information extraction from text with information extracted from graphic portions of the document.

c. Cataloging Issues: We will study the use of cataloging systems that provide users with retrieval that is customized to their needs.

d. We will hold the second of three annual symposia. In this symposium, we expect to augment the technical presentations with presentations by end-users who have had the opportunity to work with the digitized information.

During the fourth year we expect to have a fully automated conversion system operational. We will focus our efforts on completing the system and the conversion of the entire polar antarctic database, dissemination of the results, and collection of feedback from users. The final symposium will highlight the successes of the program from the research and user communities as well as identify areas for future growth and development. In the final year, we will also focus on marketing the algorithms and the testbed system to information providers for exploitation of other information collections.

IX.2.2 Testbed System

The annual milestones for the testbed system development are geared towards three goals: (a) to provide timely and accurate training data extracted from microform documents, (b) to provide researchers a testbed on which experiments can be run to evaluate existing algorithms and into which new algorithms can be integrated, and (c) to provide end users with progressively increasing functionality that they can use and comment on.

During the first year of the proposed project, the subset document database (approximately 500 documents) will be established, as a prototype, for software development and testing and training purposes. Microform documents will be scanned and converted to an image database, using current ISRI equipment augmented by microform scanners purchased with NSF funding. Bibliographic control, established by the Library of Congress will be attached to each scanned document. Also, the scientist(s) in possession of each document will be interviewed in person, for their subjective annotations and for their ideas regarding optimal modes of browsing and types of queries .

After all documents are converted, a workshop of polar antarctic scientists will be convened to examine and validate the results, and to explore retrieval procedures (current and desired). General types of retrieval questions will be noted, along with specific examples. Wherever possible, the "optimal responses" to various retrieval questions will be established, according to these scientists, who are, in fact, the domain experts for this subset database.

The development of algorithms for organization, browsing and retrieval will proceed against this prototype database during the first two years of this project.

During the proposal's second year, the collection of documents scanned in the first year will be disseminated to researchers and users as discussed in Section IX. Feedback will be solicited and incorporated into the testbed. In parallel, additional documents will be scanned, bringing the total number of documents to 10,000 by the end of the third year.

Beginning in the proposal's third year, the full English language Antarctic research literature will be converted from microform to database, and bibliographic control attached to each document as before. This (much) larger database will be used for system tuning and demonstration purposes. A joint workshop of polar antarctic scientists and digital library researchers will be held to exchange information and guide further research.

In the fourth year, a final workshop of user scientists and researchers will be convened, at which performance of the fully developed system will be considered against both the prototype and the demonstration databases. Participants will be invited to re-validate and comment on the quality and performance of retrievals, the usability of the system generally, and its likely value to them in future research.

X. COST SHARING

The 25% cost sharing requirement will be met from a combination of sources of funds. At UNLV/ISRI 50% matching funds for most senior staff salaries (not including Kanai) will be provided from State of Nevada or from private funds. The salaries of all supporting research staff, except Nagy, will will also be matched (i.e., 50-50) by funds from the State of Nevada. The additional matching funds needed for this project will be provided from UNLV Institutional funds.

SRI management has approved $350,000 (contract equivalent) in 1994 IR&D funding for research in the area of document understanding. The research program that is being conducted under this IR&D funding is directly relevant to the proposed DLI interests. Consequently, under the FAR: 31.205-18(e), SRI proposes that $300,000 of these IR&D funds be counted as cost sharing. All research results developed under these IR&D efforts will be made available to the UNLV team for integration on the testbed. Wherever appropriate, digitized data from microforms will be used during algorithm development and testing under the IR&D efforts. Further justification and documentation for the use of SRI's IR&D funds will be provided upon request by the NSF.

The 25% matching funds for both the DRI and the UNR subcontracts will be provided by State of Nevada funds. The total matching funds for this project will be $1,735,000.

XI. CONCLUSIONS

In summary, our proposal is motivated by our strong belief that as large, complete collections of significant documents become available via the medium of digital libraries, users of this information will demand complex and rich interaction with the information. Document access mechanisms will have to grow beyond keywords and full-text searches to include user-friendly mechanisms such as browsing, searching of images, and searching on the basis of abstract concepts embodied in the documents.

Authoring systems now being developed and studied will provide the appropriate tools for creating links between information elements and the ability to represent abstract concepts. Such tools, however, are most appropriate for organizing information when it is being created. It is extremely labor intensive to retroactively convert documents that already exist in hardcopy form. And unless the process of converting existing documents in a cost effective manner is developed, large fractions of the extant knowledge will not be cost effectively exploited.

Our proposal, therefore, addresses the notion of intelligent conversion of microform documents. Driven by users of scientific information, we propose researching aspects of the microform document conversion process including image preprocessing, recognition, postprocessing for extracting information, and natural language techniques for extracting information for creation of hypertext information links. Accurate characterization of existing and newly created algorithms will allow generation of a system that automatically adapts to a wide range of image quality, thereby allowing large scale conversion efforts to proceed rapidly. This research will yield processing techniques that can organize information in ways that will support the end-user in using digital information.

To make the problem tractable, we propose focusing on a particular collection of documents: the NSF Antarctic database consisting of approximately 55,000 documents that cover all published material in the field. We propose close interaction with the research community that uses these documents, including the creation of a testbed for their use, and providing mechanisms such as phased releases on which their acceptance of the techniques can be investigated.

If successful, the techniques developed could then be applied to other document collections. The algorithms and systems would be usable by information providers to convert other corpora of printed/microform documents. The research itself will add significantly to the literature in the digital libraries area as well as to the intelligent document processing literature.

We are confident that by combining the expertise of our team: the experimental research methodologies and the state-of-the-art document scanning and recognition facilities available at the ISRI at UNLV; the research capabilities in image processing, AI, natural language, hypertext, and data storage and retrieval at SRI; the end-user knowledge at the DRI; and library support for cataloging and copyright issues at UNR; we will be able to make a significant contribution within the funding and time frame of the NSF Digital Libraries Initiative. We are committed to this work as demonstrated by our significant cost sharing and look forward to its execution.

ORGANIZATIONAL ROLES

PARTICIPATING ORGANIZATIONS

Our team is strong and uniquely capable of conducting the proposed research. The Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas was established in 1990 by a grant from the United States Department of Energy. The overall mission of ISRI is to foster the improvement of automated technologies for document processing. Since 1990, ISRI has been conducting large scale experiments in document understanding using its automated experimental environment. No other research institute in the world currently has such capabilities. We have a track record in characterization of OCR technology and in investigation of the interaction between the recognition and retrieval tasks.

SRI International is the world's largest independent nonprofit research institute, chartered in the state of California, performing a broad spectrum of problem-oriented research. Groups at SRI that are involved in the proposed work have a significant track record in OCR and image understanding (e.g., for the US Postal Service), hypertext and multimedia, artificial intelligence, and distributed information storage and retrieval. Our natural language system (FASTUS), developed in 1992, was among the top two or three leaders in two recent evaluations of written language systems, and an order of magnitude faster than comparable systems.

The Desert Research Institute is the world's largest multidisciplinary organization conducting environmental research in arid lands. DRI was recently awarded a 6-year grant to study the McMurdo Dry Valleys region of Antarctica - a cold desert ecosystem - as part of the NSF's Long-Term Ecological Research (LTER) program. Over 25 senior investigators and graduate students are involved in this LTER project, making DRI home to the majority of active researchers in this area. Project staff are intimately familiar with the Antarctic literature regarding the Dry Valleys: they have studied all of it, and developed a great deal of it: they are the domain experts.

DRI originally grew out of the University of Nevada, Reno; and the two institutions maintain a close working relationship today, not only in science, but in service facilities as well, including their Libraries. For a number of years, the UNR Library has aggressively pursued electronic storage and retrieval for its holdings, which include a large collection of Government documents, on environmental matters in particular. Library staff are professionally interested in the effects of information technology on scientists' research habits and scientific productivity.

Under this program, ISRI will conduct research in OCR techniques, information retrieval, and document markup. ISRI will also be responsible for the generation of image and truthed character data for supporting the team's research activities. In cooperation with the other team members, ISRI will design, develop, document, operate, and maintain the experimental testbed. SRI will focus on preprocessing algorithms, graphics understanding techniques, and natural language techniques based on SRI's FASTUS system. DRI will provide the team access to the documents in the Polar Antarctic database and access to researchers who use the information. DRI scientists will also serve as test users of the data as it is converted and throughout the development cycle, work closely with the team in defining needs and evaluating technical approaches to meet the needs. UNR will provide bibliographical knowledge that will be required to organize the extracted information according to accepted conventions and patterns. UNR will also have the responsibility to obtain all required copyright releases and other administrative protocols for providing users controlled access to the data.

The entire team is committed to achieving the proposed goals and is enthusiastic about the program potential. Both the administration at UNLV and UNR as well as the management at SRI and DRI strongly support the proposed project.

COST SHARING

The 25% cost sharing requirement will be met from a combination of sources of funds. At UNLV/ISRI 50% matching funds for most senior staff salaries (not including Kanai) will be provided from State of Nevada or from private funds. The salaries of all supporting research staff, except Nagy, will will also be matched (i.e., 50-50) by funds from the State of Nevada. The additional matching funds needed for this project will be provided from UNLV Institutional funds.

SRI management has approved $350,000 (contract equivalent) in 1994 IR&D funding for research in the area of document understanding. The research program that is being conducted under this IR&D funding is directly relevant to the proposed DLI interests. Consequently, under the FAR: 31.205-18(e), SRI proposes that $300,000 of these IR&D funds be counted as cost sharing. All research results developed under these IR&D efforts will be made available to the UNLV team for integration on the testbed. Wherever appropriate, digitized data from microforms will be used during algorithm development and testing under the IR&D efforts. Further justification and documentation for the use of SRI's IR&D funds will be provided upon request by the NSF.

The 25% matching funds for both the DRI and the UNR subcontracts will be provided by State of Nevada funds. The total matching funds for this project will be $1,735,000.

MANAGEMENT PLAN

All development of new algorithms for pre- and post-processing will be done by the research staff at SRI and at ISRI. Comparison testing of competing technologies will be done by the technology assessment group at ISRI. The "ground-truth" test data needed will be prepared by the data preparation laboratory at ISRI. Sample queries and relevance judgment data will be provided by the DRI. Beginning in year two, continuous subjective evaluation of the prototype Digital Library system will be provided by LTER scientists at DRI. Traditional bibliographic quality control will be provided by the UNR Library. Agreements with publishers regarding copyright issues will also be obtained and managed by the UNR Library for this project.

ISRI will coordinate the efforts of all participants to insure the success of this project.

Budget removed.

BIBLIOGRAPHY

[Appelt et al., 93] Appelt, Douglas E., Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. 1993. "The SRI MUC-5 JV-FASTUS Information Extraction System." *Proceedings , Fifth Message Understanding Conference (MUC-5)*, Baltimore, Maryland, August 1993.

[Baird92] Baird, H.S. Document Image Defect Models, in H.S. Baird, H. Bunke, and K. Yamamoto (eds), Structured Document Image Analysis, pp. 546-556, Springer-Verlag: New York, 1992.

[Bototogo93] Botafogo, Rodrigo A. 1993. Cluster analysis for hypertext systems, in SIGIR '93.

[Bradford91] R. Bradford and T. Nartker, "Error Correlation in Contemporary OCR Systems," *Proc. of First International Conference on Document Analysis and Recognition*, Saint-Malo, France, 1991, pp. 516-524.

[Bradford94] Personal communication. (Mr. Bradford is Vice President of SAIC. SAIC currently holds several U.S. government contracts to convert microform collections.)

[Cardenas & McLeod90] Cardenas, A.F., and D. McLeod, eds. 1990. "Research Foundations in Object- Oriented and Semantic Database Systems," Prentice Hall Series in Data and Knowledge Base Systems, Prentice Hall, Englewood Cliffs, New Jersey.

[Catell91] Catell. 1991. Object Data Management. Object oriented and extended relational database systems, Addison Wesley Publishing Company Inc.

[Clevendon62] Cleverdon C.W., "Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," College of Aeronautics, Cranfield, England, 1962.

[Chen92] Chen, C.H., and J.L. DeCurtins, "A Segmentation-free Approach to OCR," *Proc. IEEE Workshop on Applications of Computer Vision*, pp. 190-196, 1992.

[Chen93a] Chen, C.H., and J.L. DeCurtins. 1993. "Word Recognition in a Segmentation-free Approach to OCR," Proc. 2nd Int. Conf. Document Analysis and Recognition, Ibaraki, Japan (October).

[Chen93b] Chen, C.H., and G. K. Myers. 1993. "Probabilistic Formulation for Word Recognition," Proc. Int. Workshop on Frontiers in Handwriting Recognition, pp. 379-384.

[Croft94] Croft, W.B., Harding, S., Taghva, K., and Borsack, J., "An Evaluation of Information Retrieval Accuracy with Simulated OCR Output. To appear in Proceedings, *Third Annual Symposium on Document Analysis and Information Retrieval*, April 1994, Las Vegas, NV

[Fox83] Fox, E., "Charecteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts," TR 83-561, Cornell University: Computing Science Department 1983.

[Fuller et al.93] Fuller, Michael, Eric Mackie, Ron Sacks-Davis, and Ross Wilkinson, 1993. Structured answers for a large structured document collection, in SIGIR '93.

[Harman93] Harman, Donna, "Overview of the First Text Retrieval Conference," (TREC-1), in *Proc. First Text Retrieval Conference*, pp. 1-20.

[Hearst & Plaunt93] Hearst, Marti A., and Christian Plaunt, 1993. Subtopic structuring for full-length document access, in SIGIR '93.

[Hobbs et al., 92] Hobbs, Jerry R., Douglas E. Appelt, John Bear, David Israel, and Mabry Tyson, 1992. ``FASTUS: A System for Extracting Information from Natural-Language Text'', SRI Technical Note 519, SRI International, Menlo Park, California, November 1992.

[ISRI93] Annual Report, UNLV/Information Science Research Institute, Las Vegas, NV, 1993.

[Jain93] Jain, Anil K. and Sushil Bhattacharjee "Text Segmentation Using Gabor Filters for Automatic Document Processing," Machine Vision and Applications, Vol. 5, pp. 169-184, New York, 1993.

[Jenkins94] F. Jenkins and J. Kanai, "The Use of Synthesized Images to Evaluate the Performance of Optical Character Recognition Devices and Algorithms," To appear: Proc. SPIE/IS&T's Symposium on Electronic Imaging: Science and Technology," San Jose, California, February 6-10, 1994.

[Joseph et al. 90] Joseph, J.V., S.M. Thatte, C.W. Thompson, and D.L. Wells. 1990. "Object Oriented Databases: Design and Implementation," in *Proceedings of the IEEE*, Vol. 79, No. 1, Jan. 1991, pp. 42-64.

[Kanai93] Kanai, J., T. Nartker, S. Rice, and G. Nagy, "Performance Metrics for Document Understanding Systems", *Proceedings Int. Conf. on Document Analysis and Recognition*, Tsukuba City, Japan, October 1993, page 424 .

[Kanungo93] Kanungo, Tapas, Robert M. Haralick, and Ihsin Phillips "Global and Local Document Degradation Models," Proc. Second International Conference on Document Analysis and Recognition, pp. 730-734, Tsukuba Science City, Japan, October 1993.

[Landau86] Landau. G.M.. and Uzi Vishkin. Efficient String Matching with k Mismatches. *Theoretical Computer Science.* 43 (1986) 239-249.

[Lewis92] Lewis. David D.. and Richard M. Tong. 1992. "Text Filtering in MUC-3 and MUC-4." Proceedings Fourth Message Understanding Conference (MUC-4). McLean. Virginia. June 1992. Morgan Kaufmann Publishers. Inc.. San Mateo. California. pp. 51-66.

[MacLeod90] Ian A. Macleod. "Storage and Retrieval of Structured Documents," *Information Processing & Management.* Vol. 26. No. 2. pp. 197-208. 1990.

[Mulgaonkar90] Mulgaonkar. Prasanna G. "Multiview Image Acquisition and Address-Block Location for Parcels." Proc. of United States Postal Service Advanced Technology Conference. Vol. 1, pp. 5-17, November 1990.

[Nartker92] T.A. Nartker. R. B. Bradford. and B. A. Cerny. "APreliminary Report on UNLV/GT1: A Database for Ground-Truth Testing in Document Analysis and Character Recognition." *Proc. First Symposium on Document Analysis and Information Retrieval,* Las Vegas. Nevada. March 1992.

[Nartker94a] T.A. Nartker. S.V. Rice, and J. Kanai. "OCR Accuracy: UNLV's Second Annual Test." *INFORM Magazine.* January 1994.

[Nartker94b] T.A. Nartker, "On the Need for Information Metrics," To appear in Proc. of Symposium on Electronic Imaging Science and Technology, Feb. 1994, San Jose, CA.

[Niblack93] Niblack W.. and Flickner M. "Find Me the Pictures that Look Like This: IBM's Image Query Project." Advanced Imaging, pp 32-35, April 93.

[Palumbo90] Palumbo. Paul W., Jung Soh. Sargur N. Srihari, Victor Demjanenko. and Ramalingam Sridhar "Real-Time Address Block Locations using Pipelining and Multiprocessing," ***Proc. of United States Postal Service Advanced Technology Conference,*** **Vol. 1, pp. 73-87, November 1990.**

[Pavlidis92] Pavlidis. T.,"Problems in the Recognition of Poorly Printed Text," Proc. Symposium on Document Analysis and Information Retrival, pp. 162-172. University of Nevada. Las Vegas. March 1993.

[Rice92] S.V. Rice. J. Kanai. and T.A. Nartker. "A Report on the Accuracy of OCR Devices." *Technical Report ISRI TR-92-02.* University of Nevada. Las Vegas. March 1992.

[Rice93] S.V. Rice. J. Kanai. and T.A. Nartker. "An Evaluation of OCR Accuracy." *1993 ISRI Annual Report.* University of Nevada, Las Vegas. pp. 9. 1993.

[Rumbaugh et al.91] Rumbaugh. J., M. Blaha. W. Premeriani. F. Eddy. and W. Lorensen. 1991. Object-oriented modeling and design. Prentice Hall Inc., Englewood Cliffs. New Jersey.

[Salton71] Salton. G., "Automatic indexing using bibliographic citations." Journal of Documentation. 27, (1971), pp 98-100.

[Salton93b] Salton. G., J. Allan and Chris Buckley. "Approaches to Passage Retrieval in Full Text Information Systems." in *Proc. SIGIR 93*, 16th Annual ACM SIGIR Conference. pp. 49-56.

[Salton93a] Salton. Gerard and James Allan. 1993. "Selective text utilization and text traversal." in Hypertext '93.

[Schauble93] Schauble. Peter. 1993. SPIDER: A multiuser information retrieval system for semistructured and dynamic data. in SIGIR '93.

[Shlaer & Mellor88] Shlaer. S., and S.J. Mellor. 1988. Object Oriented Systems Analysis: Modeling the world in data. Yourdin Press. Prentice Hall Inc.. Englewood Cliffs. New Jersey.

[Sparck] Sparck Jones K. and Webster. Research in Relevance Weighting, British Library Research and Development Report 5553, Computer Laboratory, University of Cambridge.

[Srihari93] Srihari. Rohini. 1993. "Intelligent Document Understanding: Understanding Photographs with Captions." Proc. of the Second International Conference on Document Analysis and Recognition, (October 20-22), 664-667.

[Sundheim92] Sundheim. Beth, ed.. *Proceedings , Fourth Message Understanding Conference (MUC-4)*, McLean, Virginia, June 1992. Distributed by Morgan Kaufmann Publishers, Inc., San Mateo, California.

[Sundhem93] Sundheim. Beth. ed., *Proceedings, Fifth Message Understanding Conference (MUC-5)*, Baltimore, Maryland, August 1993.

[Taghva93a] Kazem Taghva , Julie Borsack, Bryan Bullard, Allen Condit. Post-Editing through Approximation and Global Correction, TR 93-05, March 1993.

[Taghva94a] Kazem Taghva, Teresa Love, and Debbie Wallace, A relational Design for an OCR/IR testing environment, TR94-01, Feb. 1994.

[Taghva94b] Taghva. K., Borsack. J., Condit. A., and Erva. S., "The effects of noisy data on text retrieval systems," *Journal of American Society for Information Science* 45(1): 52-58. 1994.

[Taghva94c] Taghva. K. Borsack. J., and Condit. A., An expert system for automatically correcting OCR output. To appear in Proc. of Symposium on Electronic Imaging Science and Technology, Feb. 1994, San Jose, CA.

[Taghva94d] Taghva, K. Borsack, J., and Condit, A. Results of Applying Probabilistic IR to OCR Text. Submitted for publication.

[Ullman89] Ullman, J.D., *Principles of Database and Knowledge-Base Systems*. Computer Science Press, 1989.

[Wu91] Wu, Sun, and Udi Manber. Fast Text Searching With Errors. TR 91-11. Department of Computer Science, University of Arizona. Tucson, AZ 85721.

[Zdonik and Maier90] Zdonik, S.B., and D. Maier. 1990. "Fundamentals of Object-Oriented Databases." In *Readings on Object-Oriented Database Systems*. ed. by Stanley B. Zdonik and David Maier. Morgan Kaufmann publishers. San Mateo, California. pp. 1-32.

DATE
FILMED

6/16/94

END