Automated Metadata Extraction Services

Kurt Maly

Contact: maly@cs.odu.edu

(Collaborators: Mohammad Zubair and Steven Zeil)

Outline

- Problem and Challenges
- Approach
- Target Collections
- Architecture and Implementation
- Validation
- Conclusion

Problem & Challenges

Automated Metadata Extraction for Large, Diverse and Evolving Document Collections

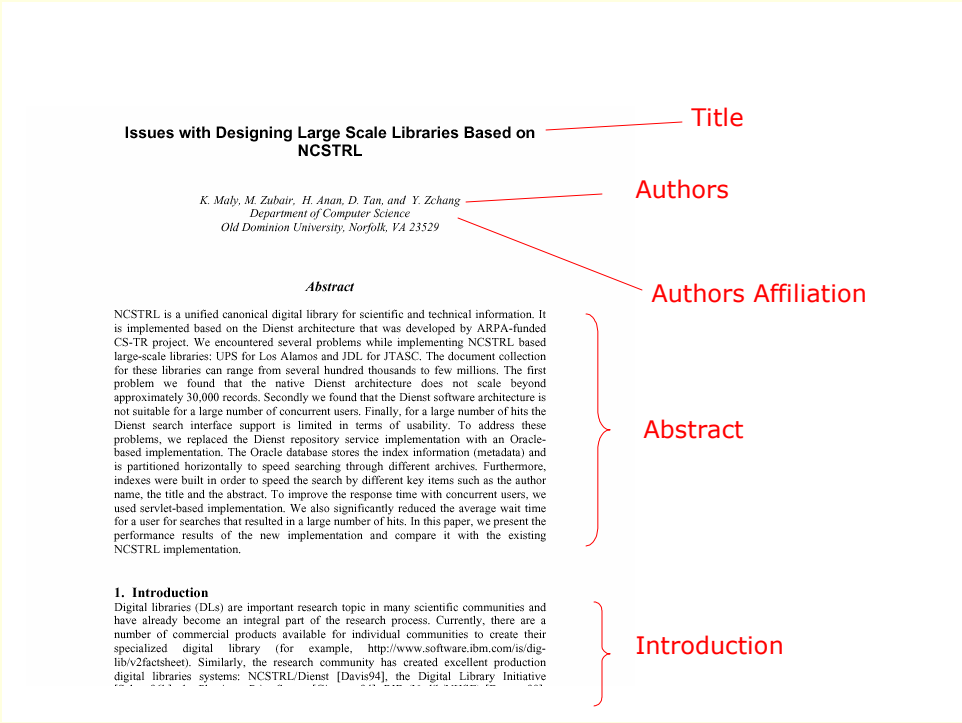
Automation ?

The traditional method of creating metadata for large collections by librarians is prohibitively expensive and time consuming. According to one estimate, it would take about 60 employee-years to create metadata for 1 million documents.

Challenges:

- Extracting metadata programmatically
- Diverse and evolving nature of a collection complicates metadata extraction

Problem & Challenges



Issues with Designing Large Scale Libraries Based on NCSTRL

Title

K. Maly, M. Zubair, H. Anan, D. Tan, and Y. Zchang
Department of Computer Science
Old Dominion University, Norfolk, VA 23529

Authors

Authors Affiliation

Abstract

NCSTRL is a unified canonical digital library for scientific and technical information. It is implemented based on the Dienst architecture that was developed by ARPA-funded CS-TR project. We encountered several problems while implementing NCSTRL based large-scale libraries: UPS for Los Alamos and JDL for JTASC. The document collection for these libraries can range from several hundred thousands to few millions. The first problem we found that the native Dienst architecture does not scale beyond approximately 30,000 records. Secondly we found that the Dienst software architecture is not suitable for a large number of concurrent users. Finally, for a large number of hits the Dienst search interface support is limited in terms of usability. To address these problems, we replaced the Dienst repository service implementation with an Oracle-based implementation. The Oracle database stores the index information (metadata) and is partitioned horizontally to speed searching through different archives. Furthermore, indexes were built in order to speed the search by different key items such as the author name, the title and the abstract. To improve the response time with concurrent users, we used servlet-based implementation. We also significantly reduced the average wait time for a user for searches that resulted in a large number of hits. In this paper, we present the performance results of the new implementation and compare it with the existing NCSTRL implementation.

Abstract

1. Introduction

Digital libraries (DLs) are important research topic in many scientific communities and have already become an integral part of the research process. Currently, there are a number of commercial products available for individual communities to create their specialized digital library (for example, http://www.software.ibm.com/is/dig-lib/v2factsheet). Similarly, the research community has created excellent production digital libraries systems: NCSTRL/Dienst [Davis94], the Digital Library Initiative

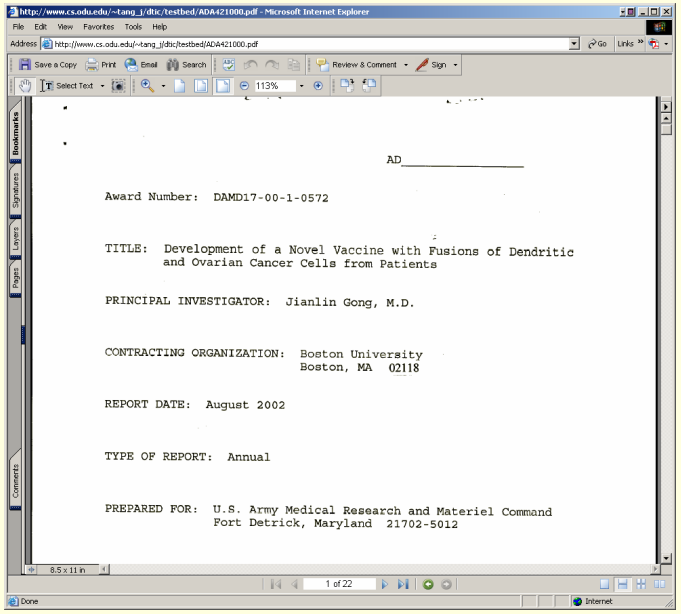
Introduction

The visual cues in the formatting of the document along with the accumulated knowledge and intelligence makes it easy for a human to identify a string as a title, author, etc. !!

What about writing a computer program to do this task automatically ?

Problem & Challenges

State-of-the-art approaches for automated metadata extraction are restricted to documents with a common layout and structure. It is relatively easy to define set of rules or train machines for a homogeneous collection.



AD_____

Award Number: DAMD17-00-1-0572

TITLE: Development of a Novel Vaccine with Fusions of Dendritic and Ovarian Cancer Cells from Patients

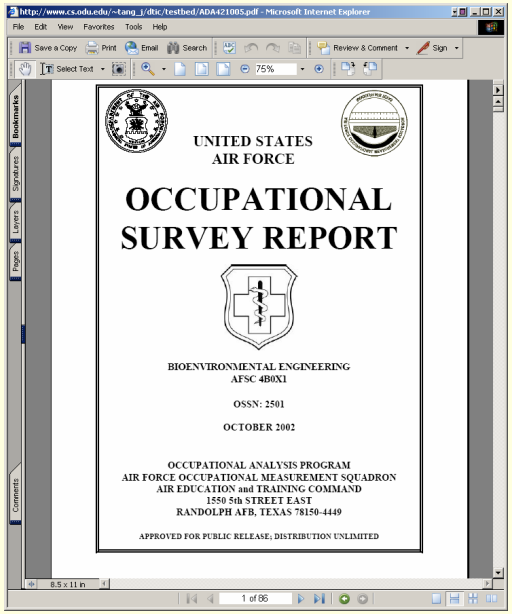
PRINCIPAL INVESTIGATOR: Jianlin Gong, M.D.

CONTRACTING ORGANIZATION: Boston University
Boston, MA 02118

REPORT DATE: August 2002

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012



UNITED STATES
AIR FORCE

OCCUPATIONAL SURVEY REPORT

BIOENVIRONMENTAL ENGINEERING
AFSC 4B0X1

OSSN: 2501

OCTOBER 2002

OCCUPATIONAL ANALYSIS PROGRAM
AIR FORCE OCCUPATIONAL MEASUREMENT SQUADRON
AIR EDUCATION and TRAINING COMMAND
1550 5th STREET EAST
RANDOLPH AFB, TEXAS 78150-4449

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

A number of federal organizations such as DTIC, GPO, and NASA manage heterogeneous collections consisting of documents with **diverse** layout and structure, where these programs do not work well.

Problem & Challenges

Evolution:

Let us assume we have developed an approach that addresses diversity present in a given collection.

What happen to the approach when collection changes with time?

What happens to the approach when the collection changes over time?

- new types of documents added
- new layouts encountered

Problem & Challenges

Robust:

A commercially viable process for metadata extraction must remain robust in the presence of external sources of error as well as in the face of the uncertainty that accompanies any attempts to automate “intelligent” behavior

Example source of error: Scanned documents with text obscured by smudges, signatures, or poor template

Approach

Existing automated metadata extraction approaches:

- Learning systems such as SVM, and HMM
  - Restricted to homogeneous collection
  - Problem with evolution: inertia to change until a significant number of examples of the new characteristics have been encountered.
- Rule-Based Systems
  - Heterogeneity can result in complex rule sets whose creation and testing can be very time-consuming. Complexity grows much more linearly in the number of rules.

Approach: Meeting the Challenges

- **Heterogeneity**
  - A new document is classified, assigning it to a group of documents of similar layout – reducing the problem to multiple homogeneous collections
  - Associated with each class of document layouts is a template, a scripted description of how to associate blocks of text in the layout with metadata fields.

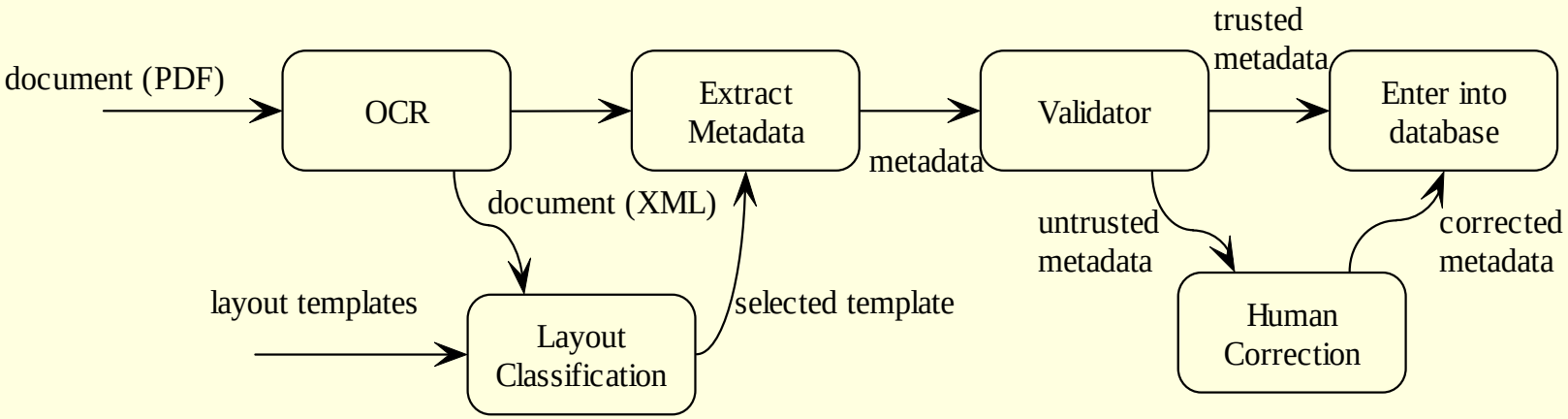
Approach: Meeting the Challenges

- **Evolution**
  - New classes of documents accommodated by writing a new template
    - templates are comparatively simple, no lengthy retraining required
    - potentially rapid response to changes in collection

Approach: Meeting the Challenges

- **Robustness**
  - Use of Validation techniques to detect extraction problems and selection of templates

Approach


document (PDF)
OCR
Extract Metadata
metadata
Validator
trusted metadata
Enter into database
document (XML)
untrusted metadata
corrected metadata
layout templates
Layout Classification
selected template
Human Correction

Target Collections

DTIC (Defense Technical Information Center) and NASA (National Aeronautics and Space Administration) Collections, Government Printing Office (GPO)

- Millions of Documents.
- Tens of thousands of new documents each year.
- Diverse: scientific articles, slides from presentations, PhD theses, (entire) conference proceedings, promotional brochures, public laws, and acts of Congress.
- Contributors: several organizations with their own in-house standards for layout and format.

Around 50% of the documents contain a Report Document Page (RDP) – a standardized form that is inserted into the document when the document is added to the collection

Sample GPO Document – With Report Document Page

Extracted Metadata

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AR-05/30		

4. Title and Subtitle
INSPECTION DEVELOPMENT FOR TITANIUM BILLET—ENGINE TITANIUM CONSORTIUM PHASE II

5. Report Date
September 2005

6. Performing Organization Code

7. Author(s)
Mike Keller[1], Thadd Patton[1], Andrei Degtyar[2], Jeff Umbach[2], Waled Hassan[3], Andy Kinney[3], Ron Roberts[4], Frank Margetan[4], and Lisa Brasche[4]

8. Performing Organization Report No.

9. Performing Organization Name and Address
[1]General Electric Company
Cincinnati, Ohio 45215

[3]Honeywell Engines, Systems & Services
Phoenix, AZ

[2]Pratt & Whitney
East Hartford, CT

[4]Iowa State University
Ames, IA

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DTFA0398FIA029

12. Sponsoring Agency Name and Address
U.S. Department of Transportation
Federal Aviation Administration
Office of Aviation Research
Washington, DC 20591

13. Type of Report and Period Covered
Final Report

14. Sponsoring Agency Code
ANE-110

15. Supplementary Notes
The FAA William J. Hughes Technical Center Technical Monitors were Rick Micklos and Cu Nguyen.

16. Abstract
The Engine Titanium Consortium (ETC) is comprised of Iowa State University; General Electric; Honeywell Engines, Systems & Services; and Pratt & Whitney. The ETC Phase I program began in 1993 with a focus on improved inspection of titanium billet used in the production of jet engines. The Phase I program completed in 1998 included the development and evaluation of two zoned approaches to billet inspection, namely, multizone and phased array inspections. The Phase II program began in 1999 and focused on further sensitivity improvements to titanium billet using the multizone approach. The goal of the Phase II effort was to achieve a #1 flat-bottom hole sensitivity for 10" diameter billet and assess the impact of attenuation compensation procedures. This report documents the results for 5", 10", and 14" diameter billets using calibration standards in a laboratory setting.

17. Key Words	18. Distribution Statement
Titanium billet, Ultrasonic inspection, Probability of detection	This document is available to the public through the National Technical Information Service (NTIS) Springfield, Virginia 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	151	

Form DOT F1700.7 (8-72) Reproduction of completed page authorized

http://128.82.7.208:9090/dtic/meta/LPS64487.xml - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://128.82.7.208:9090/dtic/meta/LPS64487.xml

```
<?xml version="1.0" encoding="windows-1252" ?>
<metadata>
  <report_num>DOT/FAA/AR-05/30</report_num>
  <government_accession_num />
  <recipient_catalog_num />
  <title>INSPECTION DEVELOPMENT FOR TITANIUM BILLET—ENGINE TITANIUM CONSORTIUM
    PHASE II</title>
  <reportdate>September 2005</reportdate>
  <performing_organization_code />
  <authors>Mike Kellen, Thadd Patton', Andrei Degtyar2, Jeff Umbach2, Waled Hassan3,
    Andy Kinney3, Ron Roberts4, Frank Margetan4, and Lisa Brasche4</authors>
  <performing_number />
  <performing_organization>3Honeywell Engines, Systems & Services 'General Electric
    Company Phoenix, AZ Cincinnati, Ohio 45215 4Iowa State University 2Pratt & Whitney
    Ames, IA East Hartford, CT</performing_organization>
  <work_unit_num />
  <contract_grant_num>DTFA0398FIA029</contract_grant_num>
  <sponsor>U.S. Department of Transportation Federal Aviation Administration Office of
    Aviation Research Washington, DC 20591</sponsor>
  <report_type_coverage>Final Report</report_type_coverage>
  <sponsor_code>ANE-110</sponsor_code>
  <notes>The FAA William J. Hughes Technical Center Technical Monitors were Rick Micklos
    and Cu Nguyen.</notes>
  <abstract>The Engine Titanium Consortium (ETC) is comprised of Iowa State University;
    General Electric; Honeywell Engines, Systems & Services; and Pratt & Whitney. The
    ETC Phase I program began in 1993 with a focus on improved inspection of titanium
    billet used in the production of jet engines. The Phase I program completed in 1998
    included the development and evaluation of two zoned approaches to billet
    inspection, namely, multizone and phased array inspections. The Phase II program
    began in 1999 and focused on further sensitivity improvements to titanium billet using
    the multizone approach. The goal of the Phase II effort was to achieve a #1 flat-
    bottom hole sensitivity for 10" diameter billet and assess the impact of attenuation
    compensation procedures. This report documents the results for 5", 10", and 14"
    diameter billets using calibration standards in a laboratory setting.</abstract>
  <keywords>Titanium billet, Ultrasonic inspection, Probability of detection</keywords>
  <dist_statement>This document is available to the public through the National Technical
    Information Service (NTIS) Springfield, Virginia 22161.</dist_statement>
  <sec_classification_report>Unclassified Form DOT F1700.7 (8-72)
    </sec_classification_report>
  <sec_classification_page>Unclassified Reproduction of completed page
    authorized</sec_classification_page>
  <num_pages>151</num_pages>
  <price />
</metadata>
```

Discussions Discussions not available on http://128.82.7.208:9090/

Done Internet

Sample DTIC Document – Without Report Document Page

AU/ACSC/138/2000-04

AIR COMMAND AND STAFF COLLEGE

AIR UNIVERSITY

IDENTIFYING AND MITIGATING THE RISKS OF COCKPIT AUTOMATION

by

Wesley A. Olson, Major, USAF

A Research Report Submitted to the Faculty

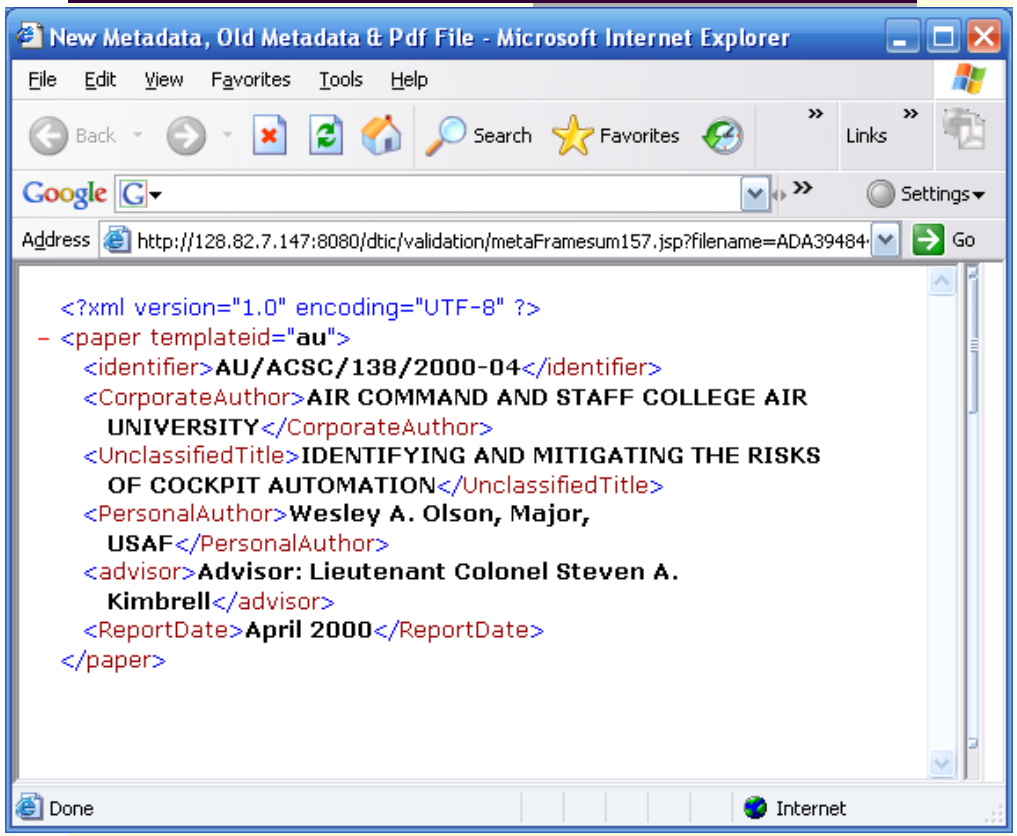
In Partial Fulfillment of the Graduation Requirements

Advisor: Lieutenant Colonel Steven A. Kimbrell

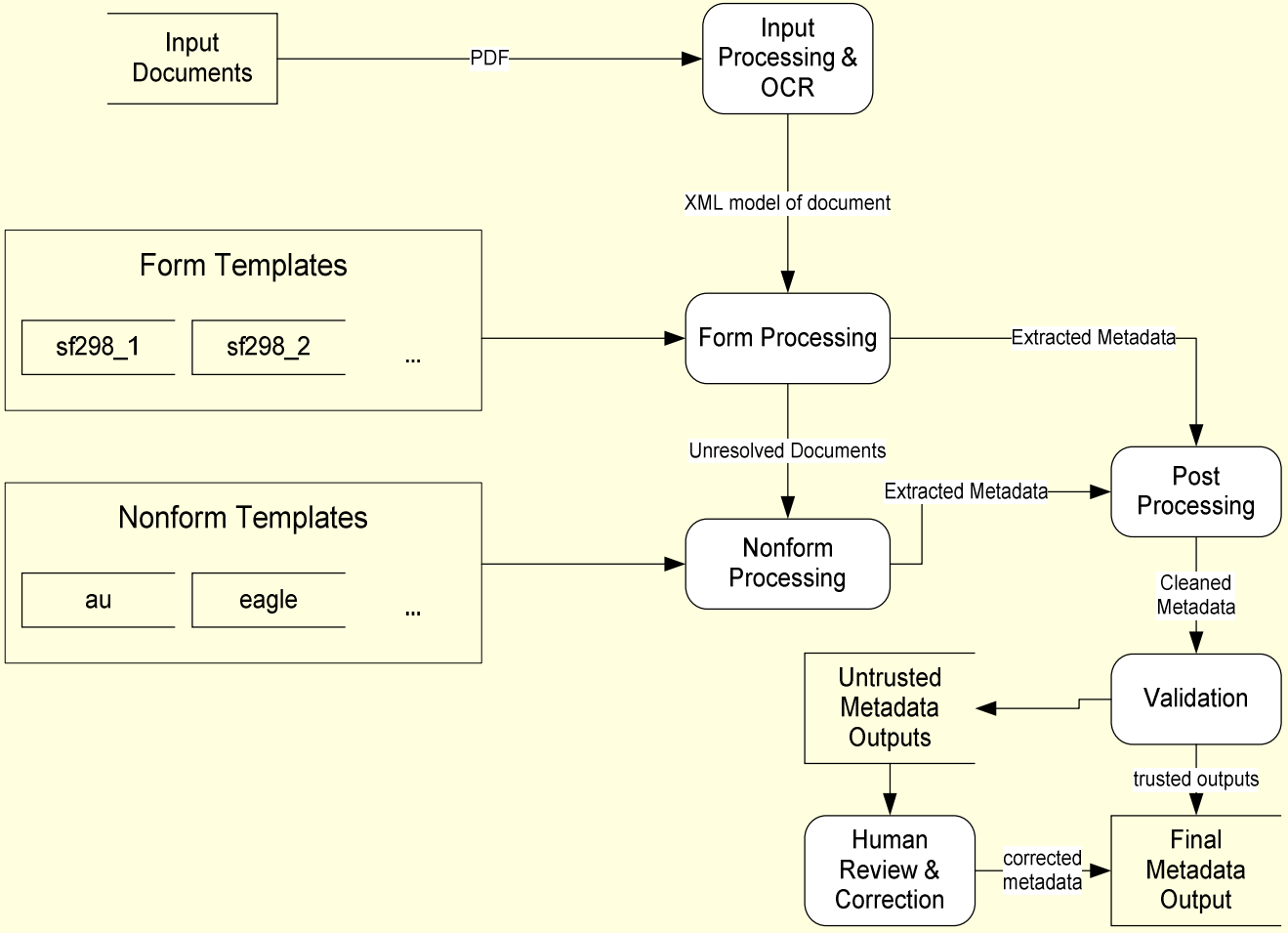
Maxwell Air Force Base, Alabama

April 2000

Extracted Metadata



Architecture & Implementation



Input Processing & OCR

- Select pages of interest from image PDF documents
  - First or last five pages of a document using the pdf toolkit (pdftk)
- Apply Off-The-Shelf OCR software
  - OmniPage Professional
- Convert OmniPage OCR output to Independent Document Model (IDM), a XML based format

Independent Document Model (IDM)

- Platform independent Document Model
- Motivation
  - Dramatic XML Schema Change between Omnipage 14 and 15
  - Tie the template engine to stable specification
  - Protects from linking directly to specific OCR product
  - Allows us to include statistics for enhanced feature usage
    - Statistics (i.e. avgDocFontSize, avgPageFontSize, wordCount, avgDocWordCount, etc..)

Documents in IDM

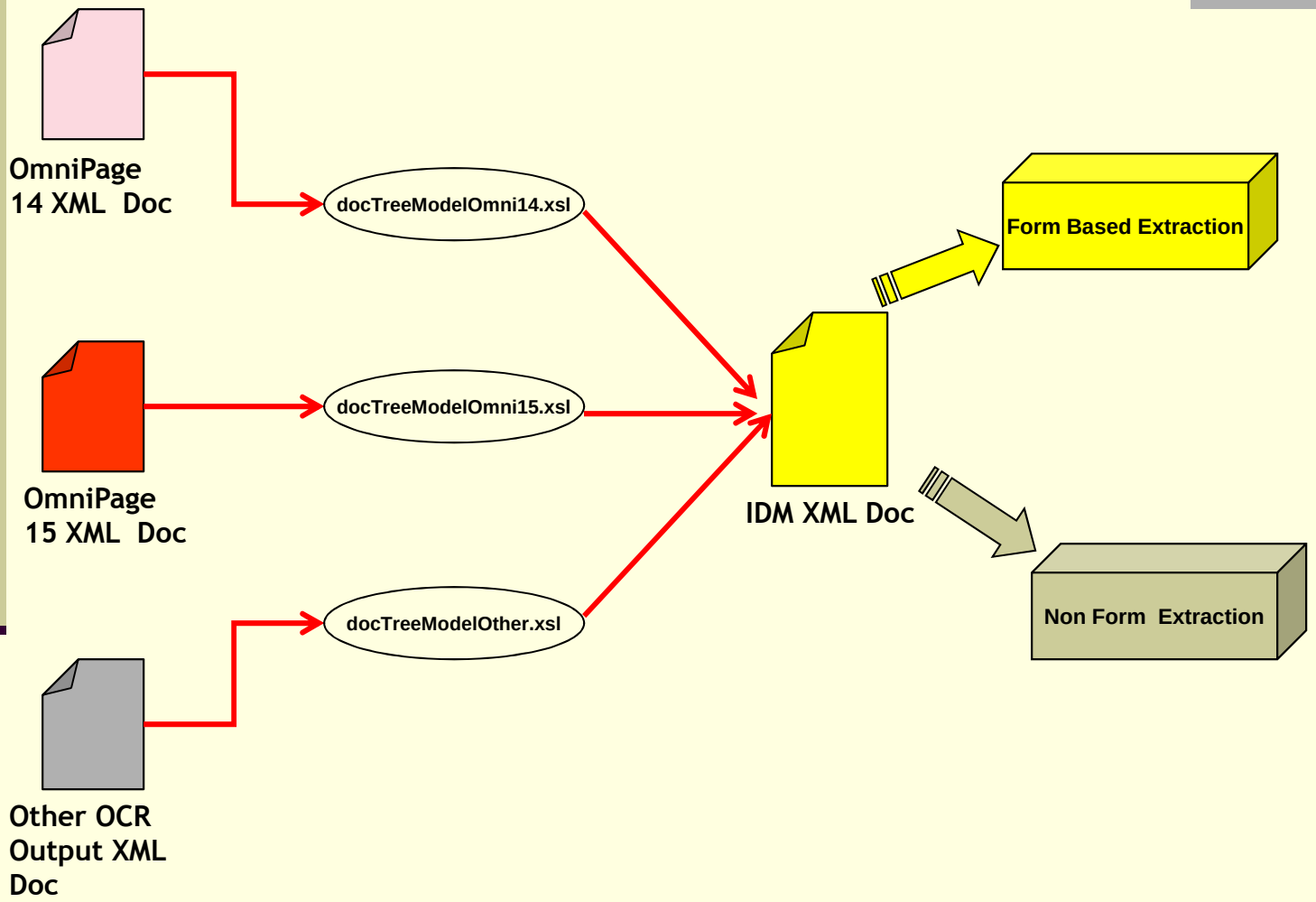
- A *document* consists of *pages*
- *pages* are divided into *regions*
- *regions* may be divided into
  - blocks of vertical whitespace
  - *paragraphs*
  - *tables*
  - *images*
- *paragraphs* are divided into *lines*
- *lines* are divided into *words*

All of these carry standard attributes for size, position, font, etc.

Generating IDM

- Use XSLT 2.0 stylesheets to transform
  - Supporting new OCR schema only requires generation of new XSLT stylesheet. -- Engine does not change

IDM Usage


OmniPage
14 XML Doc
docTreeModelOmni14.xsl
OmniPage
15 XML Doc
docTreeModelOmni15.xsl
Other OCR
Output XML
Doc
docTreeModelOther.xsl
IDM XML Doc
Form Based Extraction
Non Form Extraction

Form Processing

- Scan document to identify one of six possible forms in the DTIC collection
  - Select form template
- Form extraction engine uses the template to extract metadata from IDM based document

If the form processor fails to match any template the document moves into the non-form extraction process

Sample Form-based template fragment

```
        <field num="16->c"><line>c. THIS PAGE</line></field>
</fixed>
<extracted>
        <metadata name="ReportDate">
           <rule relation="belowof" field="1"/>
           <rule relation="aboveof" field="4|5a"/>
        </metadata>
```

The *(line)* elements in the *(field)* elements define string matching criteria.

The *(rule)* elements defined for each *(metadata)* element defines the geometric placement.

Sample Form

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.		
1. REPORT DATE (DD-MM-YYYY) 18-09-2003	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 1 April 1996 - 31 August 2003
4. TITLE AND SUBTITLE VALIDATION OF IONOSPHERIC MODELS		5a. CONTRACT NUMBER F19628-96-C-0039 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER 61102F
6. AUTHOR(S) Patricia H. Doherty Leo F. McNamara Susan H. Delay Neil J. Grossbard		5d. PROJECT NUMBER 1010 5e. TASK NUMBER IM 5f. WORK UNIT NUMBER AC
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Boston College / Institute for Scientific Research 140 Commonwealth Avenue Chestnut Hill, MA 02467-3862		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)

Sample Form (cont.)

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
Air Force Research Laboratory 29 Randolph Road Hanscom AFB, MA 01731-3010	VSBP
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-VS-TR-2003-1610

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This document represents the final report for work performed under the Boston College contract F19628-96C-0039. This contract was entitled Validation of Ionospheric Models. The objective of this contract was to obtain satellite and ground-based ionospheric measurements from a wide range of geographic locations and to utilize the resulting databases to validate the theoretical ionospheric models that are the basis of the Parameterized Real-time Ionospheric Specification Model (PRISM) and the Ionospheric Forecast Model (IFM). Thus our various efforts can be categorized as either observational databases or modeling studies.

15. SUBJECT TERMS

Ionosphere, Total Electron Content (TEC), Scintillation, Electron density, Parameterized Real-time Ionospheric Specification Model (PRISM), Ionospheric Forecast Model (IFM), Paramaterized Ionosphere Model (PIM), Global Positioning System (GPS)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John Retterer
U	U	U	SAR		19b. TELEPHONE NUMBER (Include area code) 781-377-3891

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Metadata Extracted from Sample RDP (1/3)

<metadata templateName="sf298_2">
<ReportDate>18-09-2003</ReportDate>
<DescriptiveNote>Final Report</DescriptiveNote>
<DescriptiveNote>1 April 1996 - 31 August 2003</DescriptiveNote>
<UnclassifiedTitle>VALIDATION OF IONOSPHERIC MODELS</UnclassifiedTitle>
<ContractNumber>F19628-96-C-0039</ContractNumber>
<ContractNumber></ContractNumber>
<ProgramElementNumber>61102F</ProgramElementNumber>
<PersonalAuthor>Patricia H. Doherty Leo F. McNamara
Susan H. Delay Neil J. Grossbard</PersonalAuthor>
<ProjectNumber>1010</ProjectNumber>
<TaskNumber>IM</TaskNumber>
<WorkUnitNumber>AC</WorkUnitNumber>
<CorporateAuthor>Boston College / Institute for Scientific Research 140 Commonwealth Avenue Chestnut Hill, MA 02467-3862</CorporateAuthor>

Metadata Extracted from Sample RDP (2/3)

<ReportNumber></ReportNumber>
<MonitorNameAndAddress>Air Force Research Laboratory 29 Randolph Road Hanscom AFB, MA 01731-3010</MonitorNameAndAddress>
<MonitorAcronym>VSBP</MonitorAcronym>
<MonitorSeries>AFRL-VS-TR-2003-1610</MonitorSeries>
<DistributionStatement>Approved for public release; distribution unlimited.</DistributionStatement>
<Abstract>This document represents the final report for work performed under the Boston College contract F I9628-96C-0039. This contract was entitled Validation of Ionospheric Models. The objective of this contract was to obtain satellite and ground-based ionospheric measurements from a wide range of geographic locations and to utilize the resulting databases to validate the theoretical ionospheric models that are the basis of the Parameterized Real-time Ionospheric Specification Model (PRISM) and the Ionospheric Forecast Model (IFM). Thus our various efforts can be categorized as either observational databases or modeling studies.</Abstract>

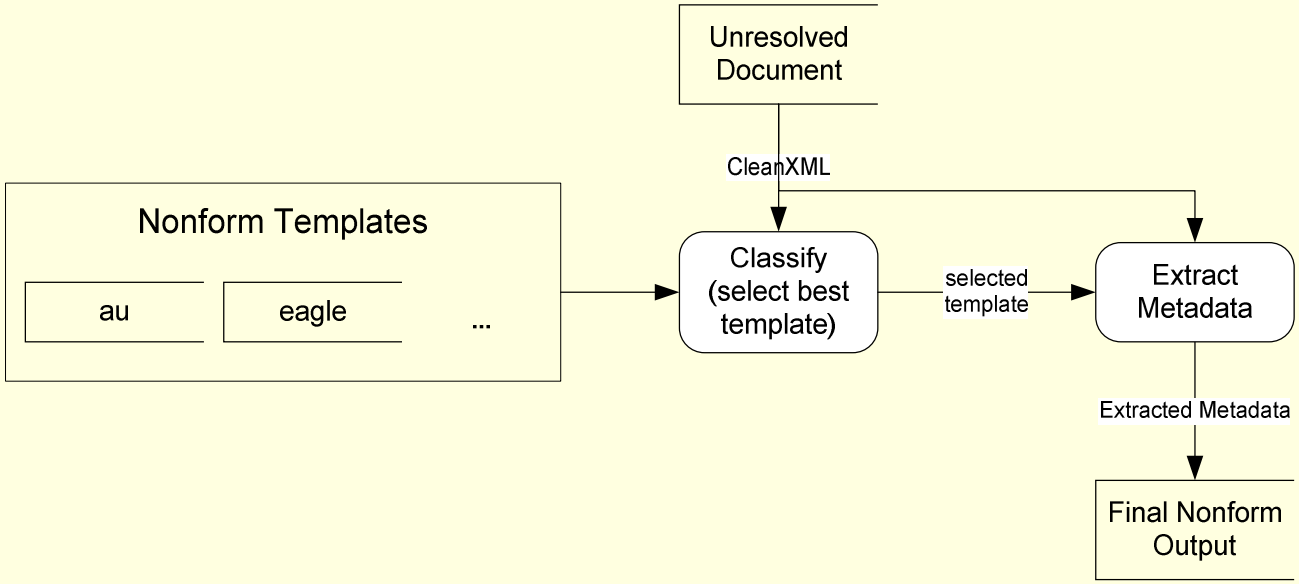
Metadata Extracted from Sample RDP (3/3)

<Identifier>Ionosphere, Total Electron Content (TEC), Scintillation, Electron density, Parameterized Real-time Ionospheric Specification Model (PRISM), Ionospheric Forecast Model (IFM), Paramaterized Ionosphere Model (PIM), Global Positioning System (GPS)</Identifier>
<ResponsiblePerson>John Retterer</ResponsiblePerson>
<Phone>781-377-3891</Phone>
<ReportClassification>U</ReportClassification>
<AbstractClassification>U</AbstractClassification>
<AbstractLimitaion>SAR</AbstractLimitaion>
</metadata>

Non-Form Processing

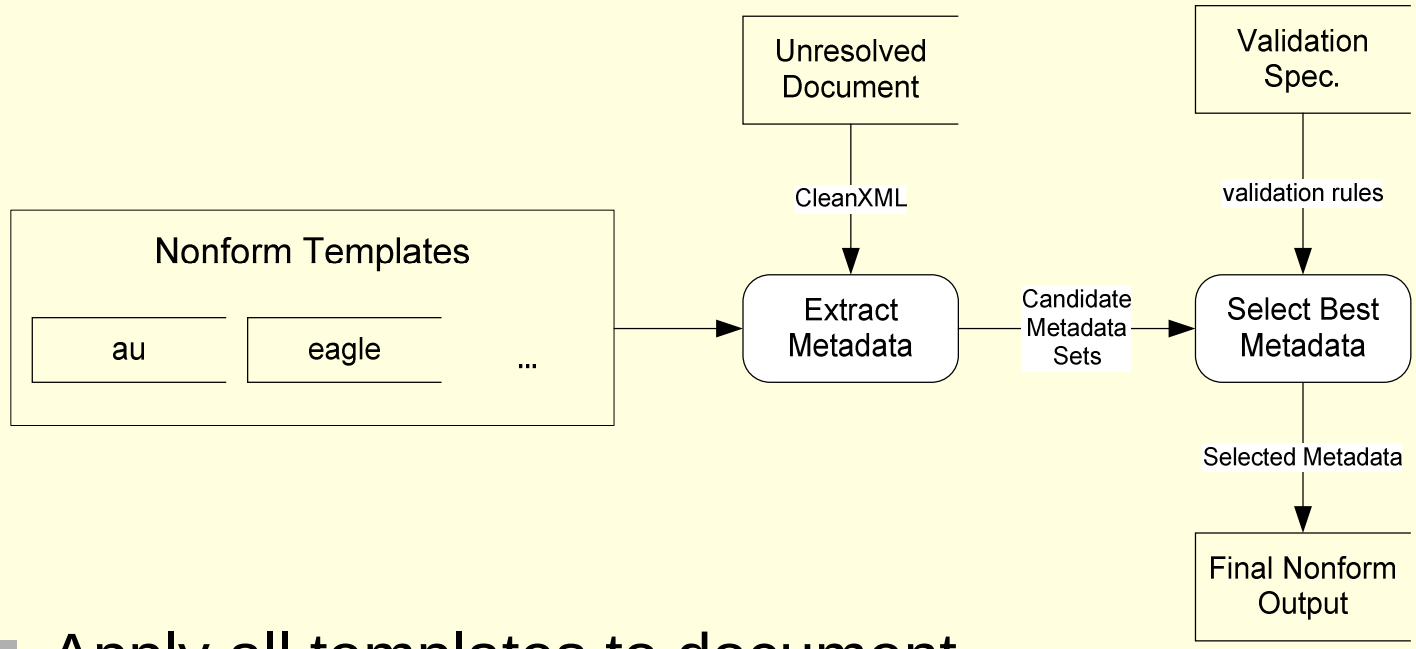
- Classification – compare document against known document layouts
  - Select template written for closest matching layout
- Apply non-form extraction engine to document and template

Classification (a priori)



- Previously, we had attempted various schemes for a priori classification
  - x-y trees, bin classification
- Still investigating some
  - image-based recognition

Post-Hoc Classification



- Apply all templates to document
  - results in multiple candidate sets of metadata
- Score each candidate using the validator
  - Select the best-scoring set

Statistical Validation of Extracted Metadata

APPROACH

Measure a relevant property, such as length of the title, and compare it to values known to be typical of documents already in that collection.

The comparison results are quantified and normalized into a confidence value for the value extracted for that metadata field.

Field	Avg.	Std. Dev.
UnclassifiedTitle	9.9	4.8
Abstract	114	58
PersonalAuthor	2.8	0.5
CorporateAuthor	7	2.3

Field Length (in words) for DTIC collection

Approach - Reference Models

We need to build models for different metadata fields to support our statistical based validation approach

- From previously extracted metadata
  - specific to document collection
- Phrase dictionaries constructed for fields with specialized vocabularies
  - e.g., author, organization
- Statistics collected
  - mean and standard deviation
  - permits detection of outputs that are significantly different from collection norms

Statistics Collected

- Field length statistics
  - title, abstract, author,..
- Phrase recurrence rates for fields with specialized vocabularies
  - author and organization
- Dictionary detection rates for words in natural language fields
  - abstract, title,.

Sample statistics collected for authors and organization names (corporate authors)

Statistics based on 800,000 human extracted metadata records

Phrase dictionaries are constructed for all phrases of length 1-4 words over a randomly selected set of 600,000 of metadata records. The remaining 200,000 are used to compute the average and standard deviation of the percentage of phrases in each field that are recorded in the phrase dictionary

Field	Phrase length	Avg.	Std. Dev.
Personal author	1	97%	11%
	2	83%	32%
	3	71%	45%
CorporateAuthor	1	100%	2%
	2	99%	6%
	3	99%	10%
	4	98%	13%

Phrase Dictionary Hit Percentage, DTIC collection

Validation Process

- Extracted outputs for fields are subjected to a variety of tests
  - Test results are *normalized* to obtain confidence value in range 0.0-1.0
- Test results for same field are combined to form field confidence
- Field confidences are combined to form overall confidence

Validation Tests

- Deterministic
  - Regular patterns such as date, report numbers
- Probabilistic
  - Length: if value of metadata is close to average -> high score
  - Vocabulary: recurrence rate according to field's phrase dictionary
  - Dictionary: detection rate of words in English dictionary

Combining results

- Validation specification describes
  - which tests to apply to which fields
  - how to combine field tests into field confidence
  - how to combine field confidences into overall confidence

Fragment for Validation Specification for DTIC Collection

```
<?xml version="1.0"?>
<val:validate collection="dtic" xmlns:val="jelly:edu.odu.cs…">
  <val:average>
    <val:field name="UnclassifiedTitle">
      <val:average>
        <val:dictionary/>
        <val:length/>
      </val:average>
    </val:field>
    <val:field name="PersonalAuthor">
      <val:min>
        <val:length/>
        <val:max>
          <val:phrases length="1"/> <val:phrases       length="2"/> <val:phrases length="3"/>
        </val:max>
      </val:min>
    </val:field>
```

Experimental Design

- How effective is post-hoc classification?
- Selected several hundred documents recently added to DTIC collection
  - Visually classified by humans,
    - comparing to 4 most common layouts from studies of earlier documents
    - discarded documents not in one of those classes
    - 167 documents remained
- Applied all templates, validated extracted metadata, selected highest confidence as the validator's choice
- Compared validator's preferred layout to human choices

Automatic vs. Human Classifications

Manually Assigned Classes	Validator au	Validator eagle	Validator rand	Validator title	Total Manual
au	86	0	0	0	86
eagle	0	8	33	4	45
rand	0	0	8	4	12
title	0	0	1	23	24

- Post-hoc classifier agreed with human on 74% of cases
- Most disagreements were due to “extra” words in extracted metadata (e.g., military ranks in author names) - highlights need for post-processing to clean up metadata
- In our simulated study of post-processing, the agreement between post-hoc classifier and human classification rose to 99%

Extracting Metadata Using Non-Form Extraction Engine

- Transform IDM based document into another XML format called CleanML, which encodes the paragraphs and lines and their corresponding features into an XML structure (this is to support our current simple non-form engine implementation)

- The non-form extraction engine also uses rule-based template extraction to locate and extract metadata. Each template contains a set of rules designed to extract metadata from a single class of similar documents.

Non-Form Template Fragment

```
<structdef pagenumber="3" templateID="arl_1">
   <CorporateAuthor>
      <begin inclusive="current">
         <stringmatch case="no" loc="beginwith">Army
             Research</stringmatch>
      </begin>
      <end inclusive="before">
         <stringmatch case="no"
            loc="beginwith">ARL</stringmatch>
      </end>
```

- Each desired metadata item is described by a rule set designating the beginning and the end of the metadata.
- The rules are limited by features detectable at the line level resolution. We hope to address this deficiency in future versions.

Non-Form Template Fragment

```
<structdef pagenumber="3" templateID="arl_1">
   <CorporateAuthor>
      <begin inclusive="current">
         <stringmatch case="no" loc="beginwith">Army
             Research</stringmatch>
      </begin>
      <end inclusive="before">
         <stringmatch case="no"
            loc="beginwith">ARL</stringmatch>
      </end>
```

Non-Form Sample (1/2)

AU/ACSC/012/1999-04

AIR COMMAND AND STAFF COLLEGE

AIR UNIVERSITY

INTEGRATING COMMERCIAL ELECTRONIC EQUIPMENT

TO IMPROVE MILITARY CAPABILITIES

by

Non-Form Sample (2/2)

by

Jeffrey A. Bohler LCDR, USN

A Research Report Submitted to the Faculty

In Partial Fulfillment of the Graduation Requirements

Advisor: CDR Albert L. St.Clair

Maxwell Air Force Base, Alabama

April 1999

Metadata Extracted From the Title Page of the Sample Document

```
<paper templateid="au">
  <identifier>AU/ACSC/012/1999-04</identifier>
  <CorporateAuthor>AIR COMMAND AND STAFF COLLEGE
    AIR UNIVERSITY</CorporateAuthor>
  <UnclassifiedTitle>INTEGRATING COMMERCIAL
    ELECTRONIC EQUIPMENT TO IMPROVE
    MILITARY CAPABILITIES
  </UnclassifiedTitle>
  <PersonalAuthor>Jeffrey A. Bohler LCDR, USN</PersonalAuthor>
  <advisor>Advisor: CDR Albert L. St.Clair</advisor>
  <ReportDate>April 1999</ReportDate>
</paper>
```

Experimental Results

- DTIC & NASA Testbed: Downloaded 9825 documents from the DTIC collection and 728 from the NASA collection.

- The internal distribution between forms and non-form documents for the collections are 94% RDP forms for DTIC and 21% RDP for NASA.

Experimental Results

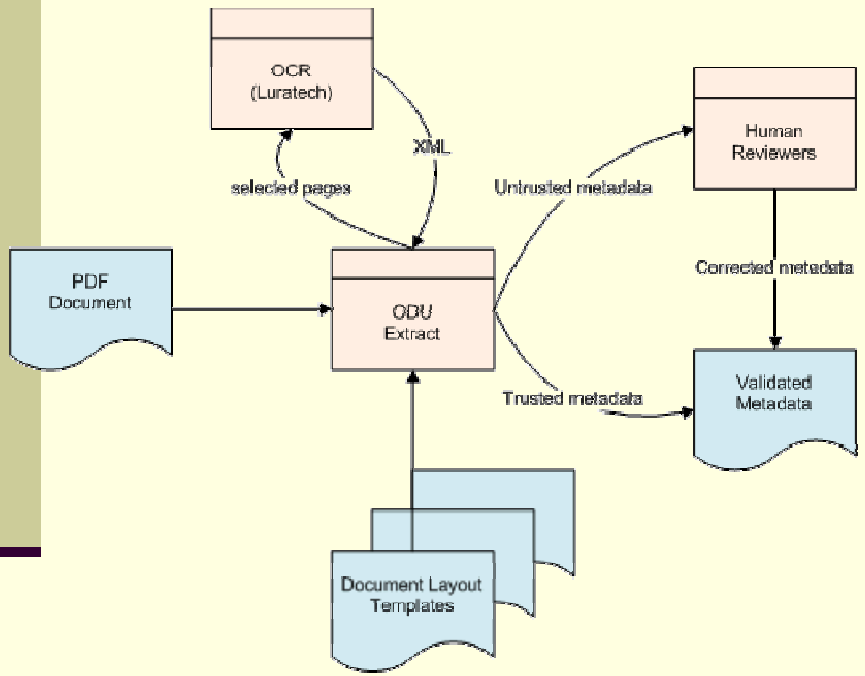
- Form based extraction
  - Wrote six form based templates
  - The overall accuracy for the for the Form based extraction was close to 99%

- Non-Form based extraction
  - Wrote 11 templates
  - The overall accuracy was 66% for DTIC and 64% for NASA. (Lower values is mostly due to the limited number of templates we were using. Assuming that we write all the necessary templates, we expect accuracy in the 85% range.

Status

- Completely Automated Software for:
  - Drop in pdf file (either text or image(scanned))
  - No OCR necessary for text pdf
  - Process and produce output metadata in XML format
  - Hooks for human intervention and correction
- Completed software for documents with RDP
- Default set of templates for:
  - RDP containing documents (about 10 templates, accuracy - 94%)
  - Non-form documents (about 100 templates, cover – 50 % of documents, accuracy – 80%)
- Statistical models of DTIC collection(800,000 documents)
  - Phrase dictionaries: personal authors, corporate authors
  - Length and English word presence for title and abstract
  - Structure of dates, report numbers

Summary – the ODU Extract System



- Documents (PDF) as input
- Invokes OCR as required
- Applies templates to find and extract metadata
- Validates its own outputs
  - Untrusted metadata outputs are diverted for human review

Summary: Elements of technology

- Partitioning the documents into classes based on visual similarity of metadata-containing pages
- Developing templates for each class, i.e., mechanisms to describe the location of metadata and how to extract them
- Develop a validation process that validates the extracted metadata against a statistical model of the collection and scores the output against the best fitting class
- Develop an automated dataflow model that starts with e-documents and ends with citations in XML or other library-specific formats; it should handle text pdfs directly and image pdf through first OCRing them
- Develop a human intervention process that will validate system suggestion of potential problems and correct them.

Summary: Advantages of technology

- Reduces time spent by humans for creating citation
  - Very significantly (greater than 90%) for documents with RDP
  - Very significantly (greater than 90%) for successful, correct extraction
  - Significantly (from 10%-90%) for successful extraction with warnings of low score
  - Significantly (from 10%-90%) for successful, partially incorrect extraction
  - None for unsuccessful extraction

Special Features of ODU System

- Post-processing
  - Author extraction conforms to DTIC cataloguing standards
    - e.g., “Capt. John James Smith, ASAF” is extracted as “Smith, John J”
  - Report dates and Descriptive notes converted as per DTIC cataloguing standards
    - will resolve ambiguous dates where possible
  - Distribution and Classification statements standardized
  - Special handling for long abstracts, metadata across multiple pages
- Dynamic Validation

  ODU Extract evaluates its own output, scores it, and flags suspicious outputs for later human inspection,
  - catches many instances of software errors, OCR errors, and human data entry errors.
  - Statistical models of DTIC collection(800,000 documents)
    - Phrase dictionaries: personal authors, corporate authors
    - Length and English word presence for title and abstract
  - Structure of dates, report numbers

Conclusions

- Automated metadata extraction can be performed effectively on a wide variety of documents
  - Coping with heterogeneous collections is a major challenge
- Much attention must be paid to “support” issues
  - validation, post-processing, etc.

Conclusions

- Creating statistical model of existing metadata is very useful tool to validate extracted metadata from new documents
- Validation can be used to classify documents and select the right template for the automated extraction process