Automated Metadata Extraction Project

Presented at the FDLP Conference,
Washington DC, Oct. 24, 2006

by
Gopi Nair

DTIC Mission

DTIC is the central scientific, research, and engineering information support activity for the Director of Defense Research and Engineering under the Office of the Secretary of Defense in executing the programs and functions of the DoD Scientific and Technical Information Program

Who We Are

Did you know that DTIC:

- Is older than the U.S. Department of Defense
- Developed one of the world’s first online bibliographic databases
- Fielded its first Web site in 1994, and currently supports more than 100 sites for the Department of Defense and military services

What We Do

- Leverage the multi-billion dollar investment in DoD research and engineering
- Prevent unnecessary or redundant research
- Get scientific and technical information into the hands of the “right” people in the defense community
- Enable the conversion of completed research into the production of mature technology

DTIC Functions

- Single point of access for Defense Acquisition, Scientific and Technical Information (STI)
- Centralized collection and secondary dissemination of STI
- Control access to information products and services
- Balance the scientific community's need for open access to information against DoD's need for limitations on access
- Manage the DoD/DTIC Information Analysis Centers
- Host Web sites for major components of DoD
- Focal point for OSD policy relating to STI

Metadata Extraction- Overview

- DTIC funded the software development effort with Old Dominion University (ODU) Digital Library Research Group in FY 04
- What is Automatic Extraction of Metadata?
  - Software that can identify and extract metadata such as Title, Personal Author, Corporate Author, Report Date, Distribution Limitations, Abstract, from an electronic document with minimal or no human intervention for the citation creation
- Benefits:
  - Citation creation is a labor- intensive process
  - Automating that process reduces operating cost
- NASA joined DTIC on this effort in FY 06

Motivation

- Metadata enhances the value of a document collection
  - Using metadata helps resource discovery
    - Save about $8,200 per employee for a company to use metadata in its intranet to reduce employee time for searching, verifying and organizing the files (estimation made by Mike Doane on DCMI 2003 workshop)

- Manual metadata extraction is costly and time-consuming
  - It would take about 60 employee-years to create metadata for one million documents (estimation made by Lou Rosenfeld on DCMI 2003 workshop).
  - Automatic extraction tools are essential to reduce cost in metadata creation as well as for rapid dissemination of content
    - OCR is not sufficient for making “legacy” documents searchable

Various Methods for Metadata Extraction

- ODU evaluated different methods to extract metadata from DTIC documents
- Machine Learning Approach
  - Support Vector Machines (SVM)
  - Hidden Markov Model (HMM)
- Template Approach - Rule-based approach
  - Using rules to specify how to extract metadata

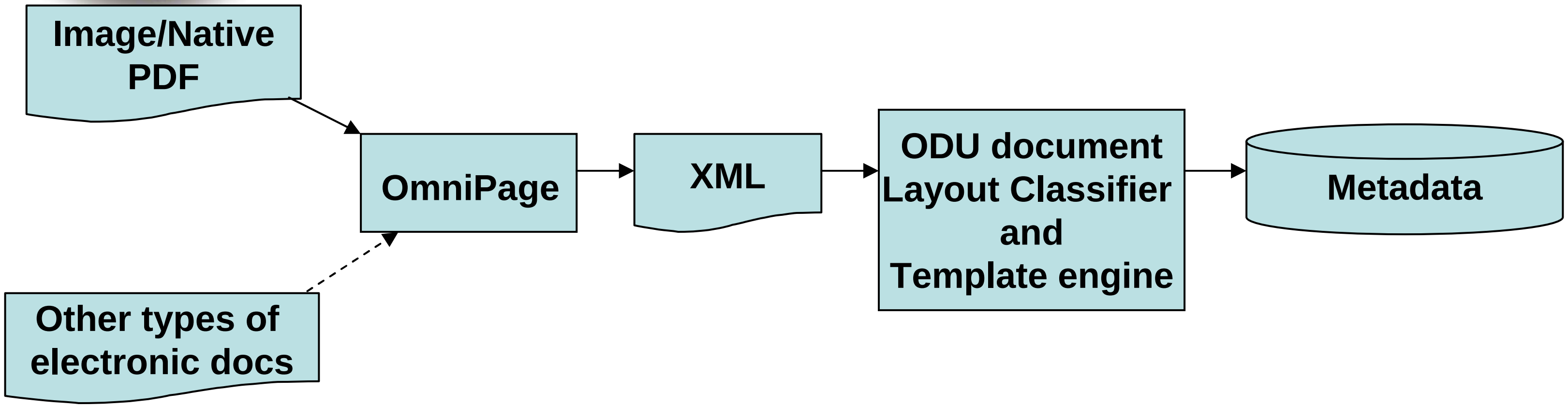
Methods Comparison

- Machine-Learning Approach
  - Good adaptability, but it has to be trained from samples – very time consuming
  - Performance degrades with increasing heterogeneity
  - Difficult to add new fields to be extracted
  - Difficult to select the right features for training
- Template Approach - Rule-based
  - No need for training from samples
  - Can extract different metadata from different documents
  - Rules can be written by non-technical people

Template Approach

- Template is an XML file describing documents with similar layout
- Uses rules to define how to extract metadata for that layout
- Documents grouped into classes based on the layout, and Template developed for each class
  - Separate Template for documents with RDP (Report Documentation Page)
- Automatic Switching Software is being developed
  - No need for manually grouping documents

Metadata Extraction Process


Image/Native PDF
Other types of electronic docs
OmniPage
XML
ODU document Layout Classifier and Template engine
Metadata

Step 1-- Convert electronic documents into XML format using OmniPage

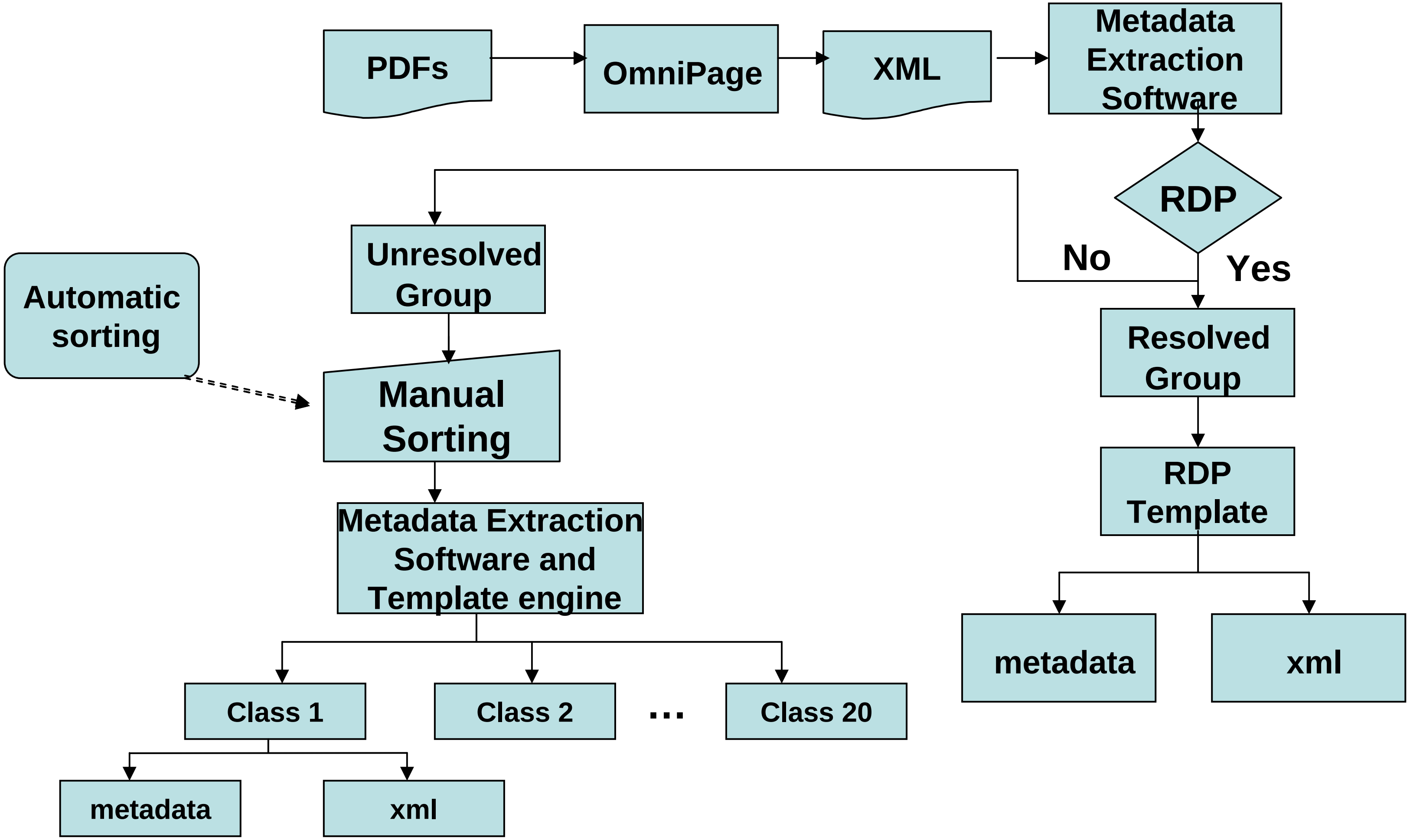
For PDF images files and Native PDF documents, OmniPage OCRs then converts into XML

OmniPage converts other types of electronic documents into XML

Step 2: Layout is recognized and grouped to select template

Step 3: Metadata is extracted from XML files using Templates

Metadata Extraction Prototype



PDFs

OmniPage

XML

Metadata Extraction Software

RDP

No

Yes

Unresolved Group

Automatic sorting

Manual Sorting

Metadata Extraction Software and Template engine

Class 1

Class 2

…

Class 20

metadata

xml

Resolved Group

RDP Template

metadata

xml

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE 7 January 2005	2. REPORT TYPE Report of Test Results	3. DATES COVERED 15-22 November 2003

4. TITLE AND SUBTITLE

USS JOHN F. KENNEDY (CV-67) Precision Approach and Landing System Verification, Final Report

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Brandon L. Jones

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Air Warfare Center Aircraft Division
22347 Cedar Point Road, Unit #6
Patuxent River, Maryland 20670-1161

8. PERFORMING ORGANIZATION REPORT NUMBER

NAWCADPAX/RTR-2004/9

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Air Systems Command (PMA-251)
47123 Buse Road Unit IPT
Patuxent River, Maryland 20670-1547

10. SPONSOR/MONITOR'S ACRONYM(S)

PMA 251 (ALRE)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution authorized to U.S. Government agencies and their contractors only: Critical Technology; January 2005. Other requests for this document shall be referred to the Naval Air Systems Command, 47123 Buse Road, Patuxent River, Maryland 20670-1547.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Flight tests were conducted aboard USS JOHN F. KENNEDY from 15 to 22 November 2003 with F/A-18A and F/A-18F aircraft to verify the AN/SPN-46(V)3 for Mode I/IA/II/IID/III operations and the AN/SPN-41 for operations as an Instrument Carrier Landing System (ICLS) and AN/SPN-46(V)3 monitor. A total of 93 Mode I approaches was completed during 8 flight periods. USS JOHN F. KENNEDY AN/SPN-46(V)3 performance is satisfactory for Mode I/IA/II/IID/III operations. AN/SPN-41 performance is satisfactory as an ICLS and AN/SPN-46(V)3 monitor. This report provides the results of the data analysis and issues final clearances.

15. SUBJECT TERMS

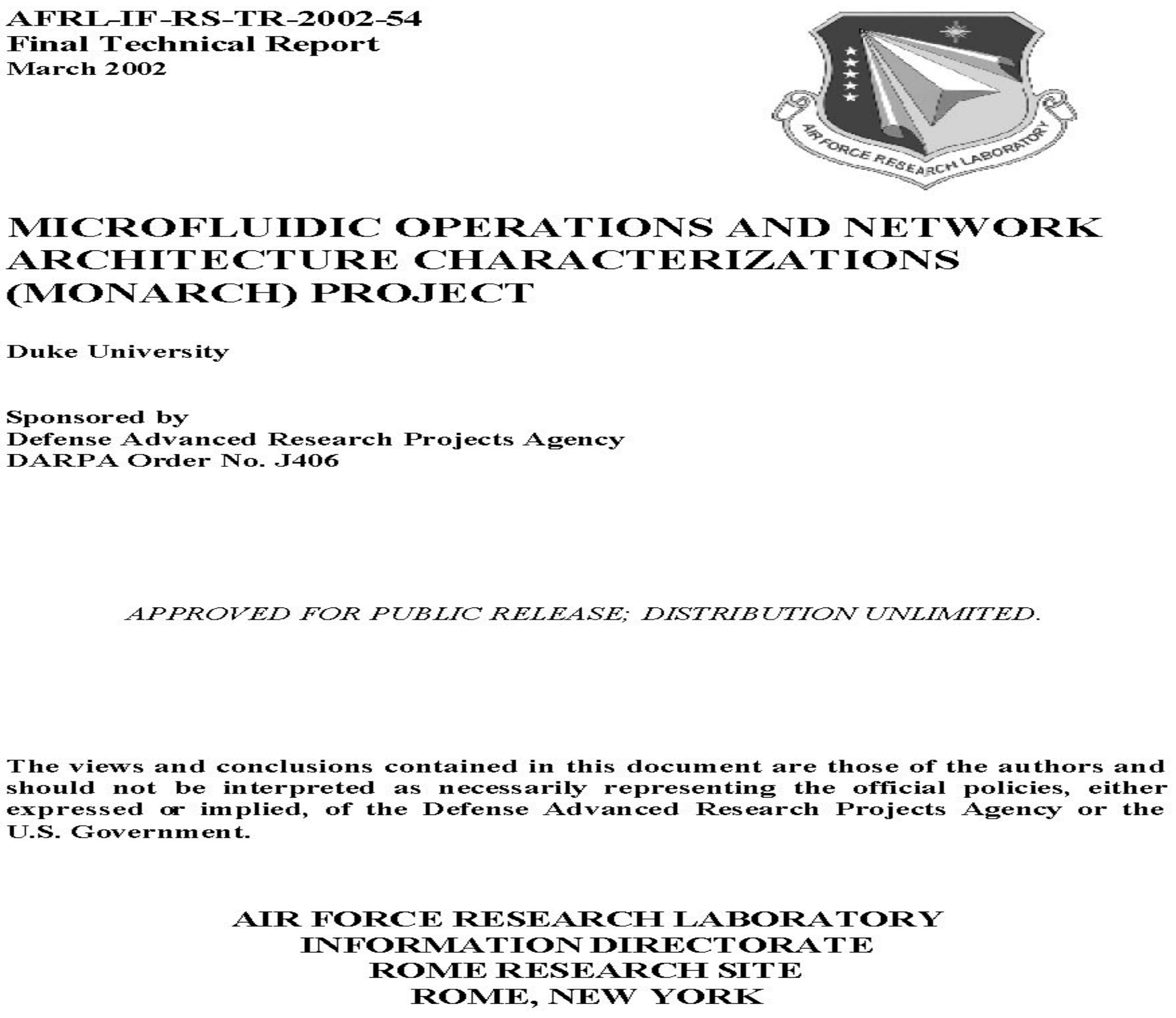
USS JOHN F. KENNEDY CV-67 Precision Approach and Landing System (PALS) Instrument Carrier Landing System (ICLS)
F/A-18A F/A-18F AN/SPN-41 AN/SPN-46(V)3

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Brandon L. Jones
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified	SAR	40	19b. TELEPHONE NUMBER (include area code) (301) 342-0775

METADATA GENERATED FROM RDP

<name>1. REPORT DATE</name>
<value>7 January 2005</value>
<name>2. REPORT TYPE</name>
<value>Report of Test Results</value>
<name>3. DATES COVERED</name>
<value>15-22 November 2003</value>
<name>4. TITLE AND SUBTITLE</name>
<value>USS JOHN F. KENNEDY (CV-67) Precision Approach and Landing System Verification, Final Report</value>
<name>6. AUTHOR(S)</name>
<value>Brandon L. Jones</value>
<name>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</name>
<value>Naval Air Warfare Center Aircraft Division 22347 Cedar Point Road, Unit #6 Patuxent River, Maryland 20670-1161</value>
<name>8. PERFORMING ORGANIZATION REPORT NUMBER</name>
<value>NAWCADPAX/RTR-2004/9</value>
<name>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</name>
<value>Naval Air Systems Command (PMA-251) 47123 Buse Road Unit IPT Patuxent River, Maryland 20670-1547</value>
<name>10. SPONSOR/MONITOR'S ACRONYM(S)</name>
<value>PMA 251 (ALRE)</value>
<name>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</name>
<value>January 2005. Other requests for this Maryland 20670-1547.</value>
<name>12. DISTRIBUTION/AVAILABILITY STATEMENT</name>
<value>Distribution authorized to U.S. Government agencies and their contractors only: Critical Technology; document shall be referred to the Naval Air Systems Command, 47123 Buse Road, Patuxent River,</value>
<name>14. ABSTRACT</name>
<value>Flight tests were conducted aboard USS JOHN F. KENNEDY from 15 to 22 November 2003 with F/A-I8A and F/A-I 8F aircraft to verify the AN/SPN-46(V)3 for Mode I/IA/WIID/Ill operations and the AN/SPN-41 for operations as an Instrument Carrier Landing System (ICLS) and AN/SPN-46(V)3 monitor. A total of 93 Mode I approaches was completed during 8 flight periods. USS JOHN F. KENNEDY AN/SPN-46(V)3 performance is satisfactory for Mode I/IA/WIID/III operations. AN/SPN-41 performance is satisfactory as an ICLS and AN/SPN-46(V)3 monitor. This report provides the results of the data analysis and issues final clearances.</value>
<name>15. SUBJECT TERMS</name>
<name>USS JOHN F. KENNEDY</name>
<value>CV-67 Precision Approach and Landing System (PALS) Instrument Carrier Landing System (ICLS) RA-18A F/A-18F AN/SPN-41 AN/SPN-46(V)3</value>
<name>16. SECURITY CLASSIFICATION OF:</name>
<name>17. LIMITATION OF ABSTRACT</name>
<value>SAR</value>
<name>18. NUMBER OF PAGES</name>
<value>40</value>
<name>19a. NAME OF RESPONSIBLE PERSON</name>
<value>Brandon L. Jones</value>
<name>16a. REPORT</name>
<value>Unclassified</value>
<name>16b. ABSTRACT</name>
<value>Unclassified</value>
<name>16c. THIS PAGE</name>
<value>Unclassified</value>
<name>19b. TELEPHONE NUMBER (include area</name>
<value>code) (301) 342-0775 Scandeld tromp 295 (Rev. a-Ya)</value>

TITLE PAGE OF AFRL DOCUMENT- Different Layout



AFRL-IF-RS-TR-2002-54
Final Technical Report
March 2002

MICROFLUIDIC OPERATIONS AND NETWORK ARCHITECTURE CHARACTERIZATIONS (MONARCH) PROJECT

Duke University

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. J406

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

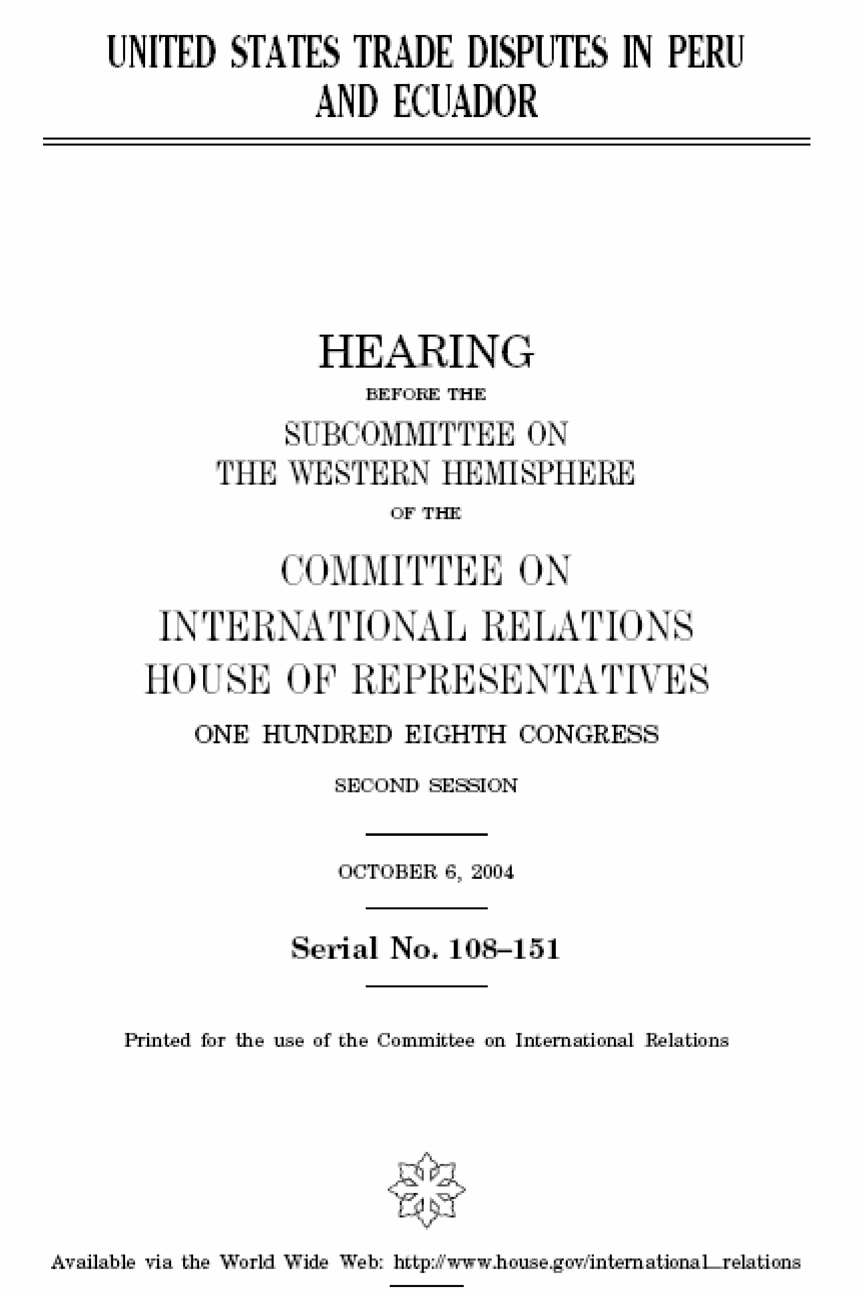
AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

Metadata Extracted From the Title Page of the AFRL Document

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<paper>
<metadata>

<Report Number>AFRL-IF-RS-TR-2002-54</Report Number>

<Descriptive Note>Final Technical Report</Descriptive Note>

<Report Date>March 2002</Report Date>

<Unclassified Title>MICROFLUIDIC OPERATIONS AND NETWORK ARCHITECTURE
CHARACTERIZATIONS (MONARCH ) PROJECT</Unclassified Title>

<Corporate Author>Duke University</Corporate Author>

<Contract Number>DARPA Order No. J406</Contract Number>

<Distribution Statement>APPROVED FOR PUBLIC RELEASE; DISTRIBUTION
UNLIMITED</Distribution Statement>

</metadata>
</paper>
```

Sample GPO Document - Without Technical Report Document Page

Extracted Metadata



UNITED STATES TRADE DISPUTES IN PERU AND ECUADOR

HEARING

BEFORE THE

SUBCOMMITTEE ON THE WESTERN HEMISPHERE

OF THE

COMMITTEE ON INTERNATIONAL RELATIONS HOUSE OF REPRESENTATIVES

ONE HUNDRED EIGHTH CONGRESS

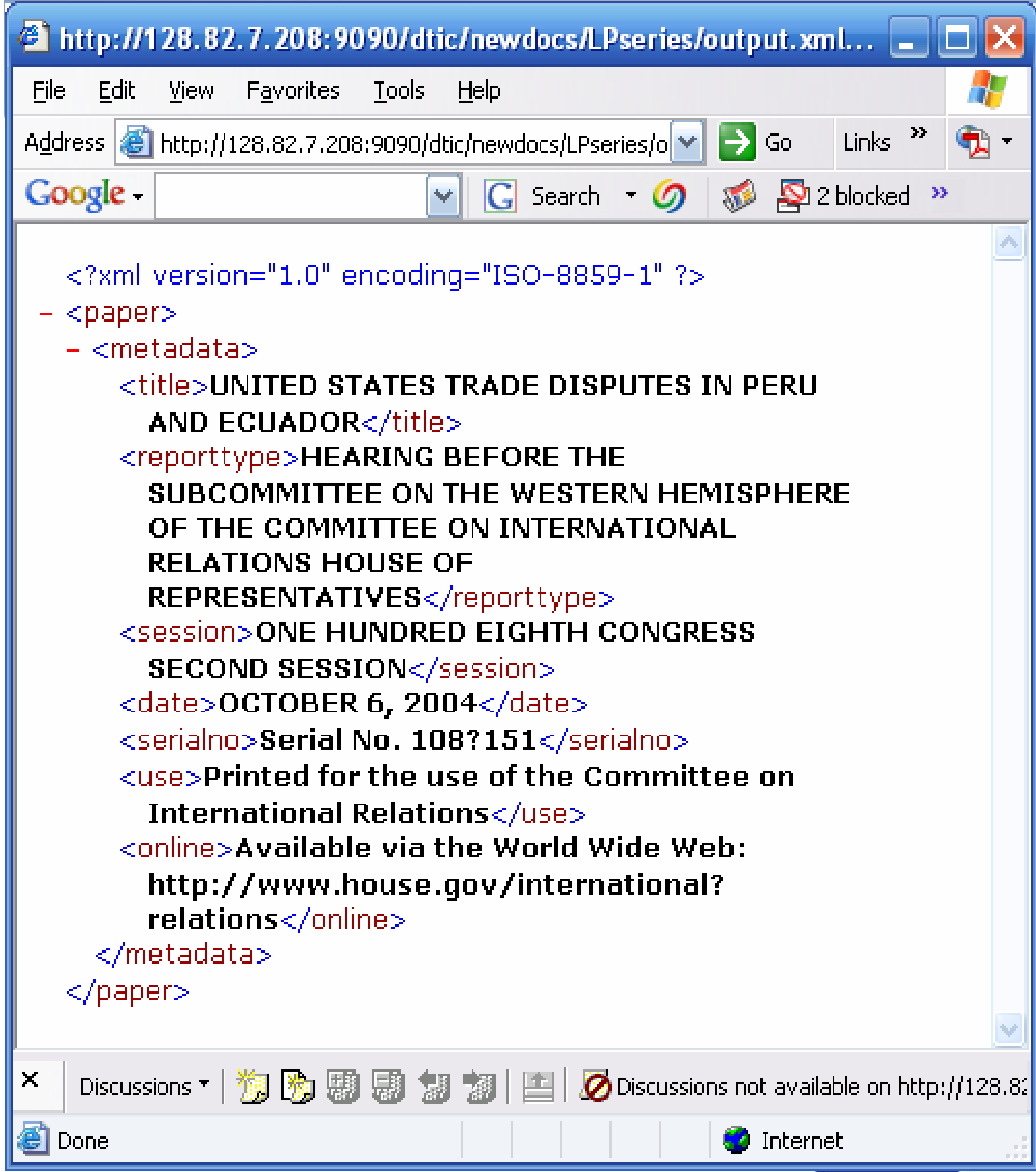
SECOND SESSION

OCTOBER 6, 2004

Serial No. 108–151

Printed for the use of the Committee on International Relations

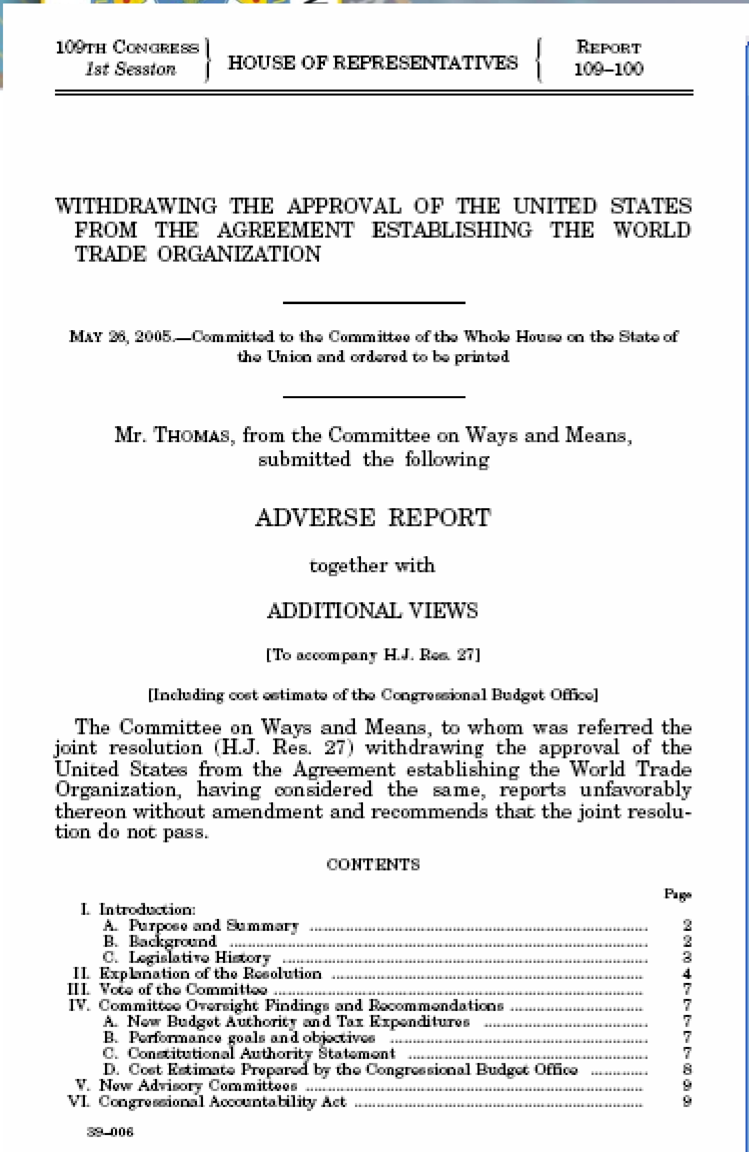
Available via the World Wide Web: http://www.house.gov/international_relations



http://128.82.7.208:9090/dtic/newdocs/LPseries/output.xml...
File
Edit
View
Favorites
Tools
Help
Address
http://128.82.7.208:9090/dtic/newdocs/LPseries/o
Go
Links
Google
Search
2 blocked
Discussions
Discussions not available on http://128.82
Done
Internet

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <paper>
  - <metadata>
      <title>UNITED STATES TRADE DISPUTES IN PERU
        AND ECUADOR</title>
      <reporttype>HEARING BEFORE THE
        SUBCOMMITTEE ON THE WESTERN HEMISPHERE
        OF THE COMMITTEE ON INTERNATIONAL
        RELATIONS HOUSE OF
        REPRESENTATIVES</reporttype>
      <session>ONE HUNDRED EIGHTH CONGRESS
        SECOND SESSION</session>
      <date>OCTOBER 6, 2004</date>
      <serialno>Serial No. 108?151</serialno>
      <use>Printed for the use of the Committee on
        International Relations</use>
      <online>Available via the World Wide Web:
        http://www.house.gov/international?
        relations</online>
    </metadata>
  </paper>
```

Sample GPO Document - Without Technical Report Document Page



109TH CONGRESS
1st Session

HOUSE OF REPRESENTATIVES

REPORT
109–100

WITHDRAWING THE APPROVAL OF THE UNITED STATES FROM THE AGREEMENT ESTABLISHING THE WORLD TRADE ORGANIZATION

MAY 26, 2005.—Committed to the Committee of the Whole House on the State of the Union and ordered to be printed

Mr. THOMAS, from the Committee on Ways and Means, submitted the following

ADVERSE REPORT

together with

ADDITIONAL VIEWS

[To accompany H.J. Res. 27]

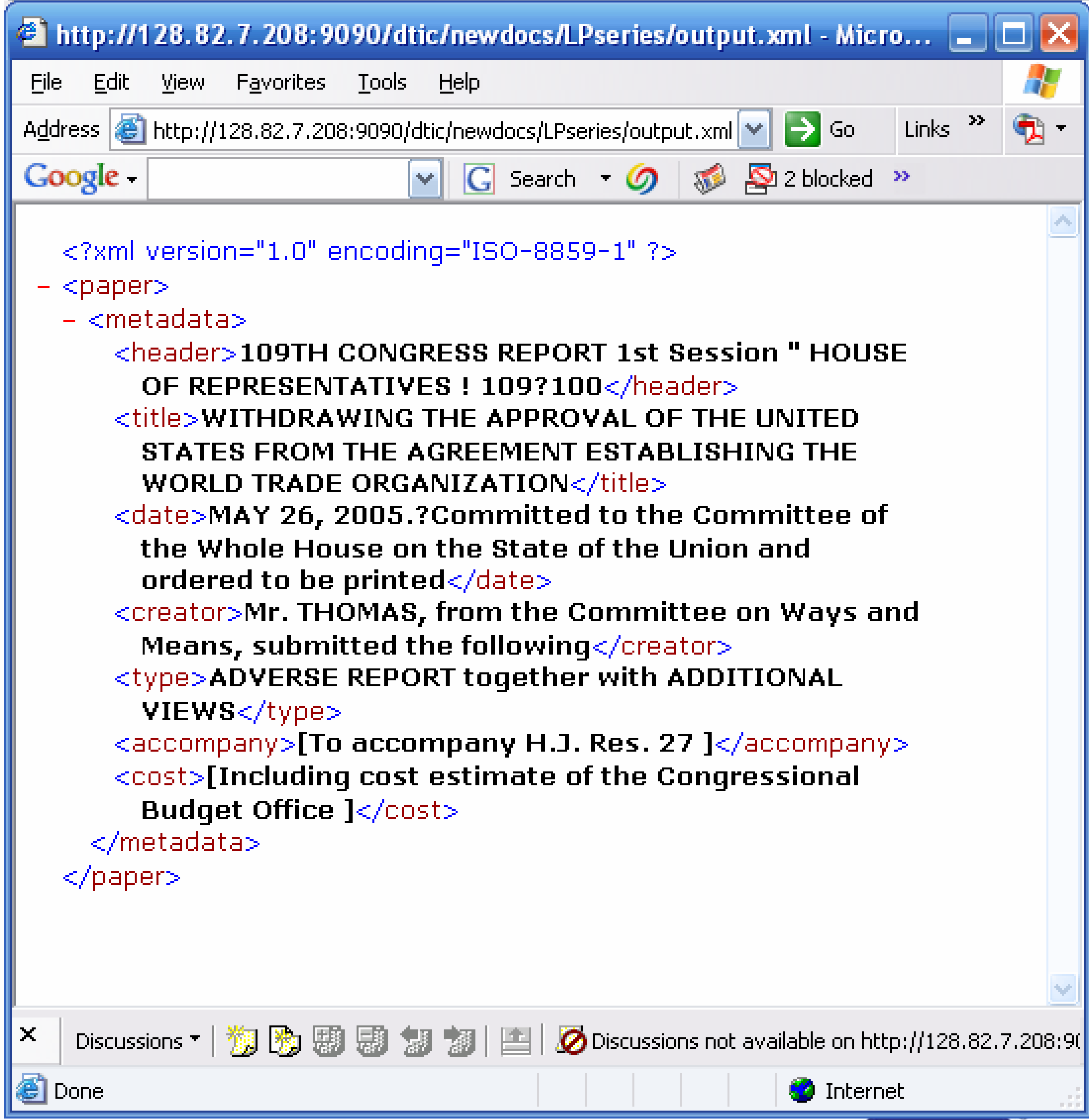
[Including cost estimate of the Congressional Budget Office]

The Committee on Ways and Means, to whom was referred the joint resolution (H.J. Res. 27) withdrawing the approval of the United States from the Agreement establishing the World Trade Organization, having considered the same, reports unfavorably thereon without amendment and recommends that the joint resolution do not pass.

CONTENTS

39–006

Extracted Metadata

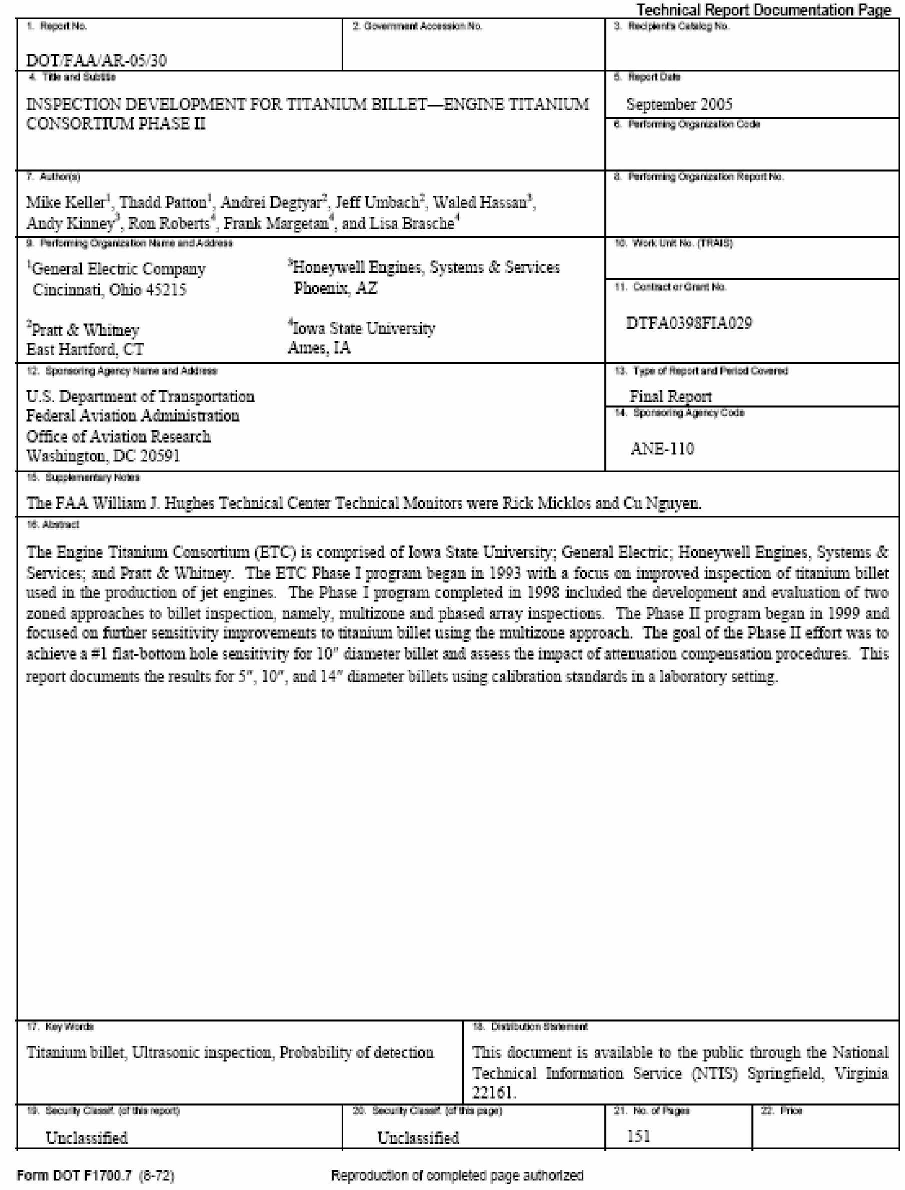

http://128.82.7.208:9090/dtic/newdocs/LPseries/output.xml - Micro...
File
Edit
View
Favorites
Tools
Help
Address
http://128.82.7.208:9090/dtic/newdocs/LPseries/output.xml
Go
Links
Google
Search
2 blocked
Discussions
Discussions not available on http://128.82.7.208:9(
Done
Internet

```
  <?xml version="1.0" encoding="ISO-8859-1" ?>
- <paper>
  - <metadata>
      <header>109TH CONGRESS REPORT 1st Session " HOUSE
        OF REPRESENTATIVES ! 109?100</header>
      <title>WITHDRAWING THE APPROVAL OF THE UNITED
        STATES FROM THE AGREEMENT ESTABLISHING THE
        WORLD TRADE ORGANIZATION</title>
      <date>MAY 26, 2005.?Committed to the Committee of
        the Whole House on the State of the Union and
        ordered to be printed</date>
      <creator>Mr. THOMAS, from the Committee on Ways and
        Means, submitted the following</creator>
      <type>ADVERSE REPORT together with ADDITIONAL
        VIEWS</type>
      <accompany>[To accompany H.J. Res. 27 ]</accompany>
      <cost>[Including cost estimate of the Congressional
        Budget Office ]</cost>
    </metadata>
  </paper>
```

Information for the Defense Community

Sample GPO Document – With Technical Report Document Page

Extracted Metadata



Technical Report Documentation Page

1. Report No. DOT/FAA/AR-05/30	2. Government Accession No.	3. Recipient's Catalog No.
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 [2]Pratt & Whitney East Hartford, CT	[3]Honeywell Engines, Systems & Services Phoenix, AZ [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

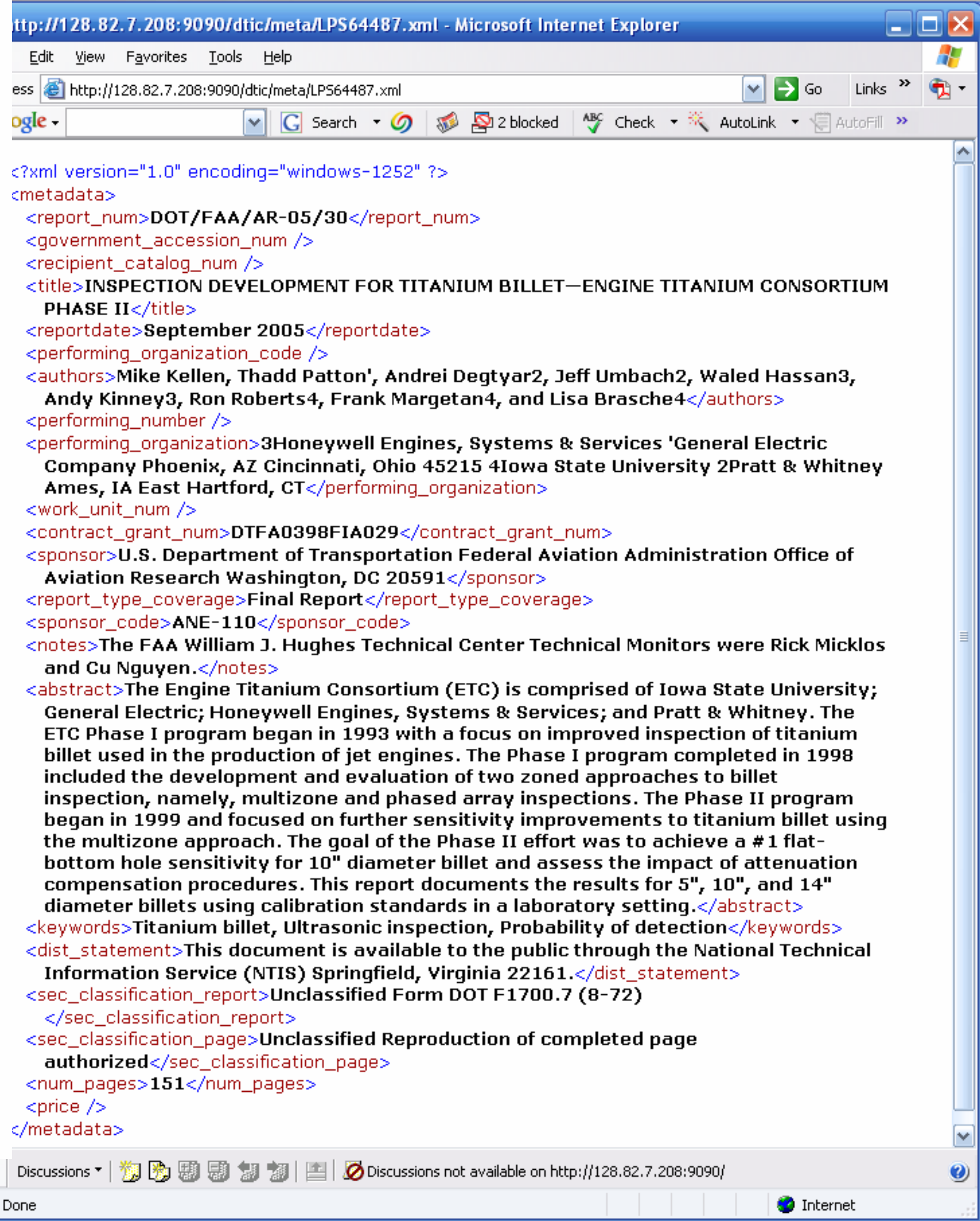
The FAA William J. Hughes Technical Center Technical Monitors were Rick Micklos and Cu Nguyen.

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 Titanium billet, Ultrasonic inspection, Probability of detection		18. Distribution Statement This document is available to the public through the National Technical Information Service (NTIS) Springfield, Virginia 22161.	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 151	22. Price

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



ttp://128.82.7.208:9090/dtic/meta/LPS64487.xml - Microsoft Internet Explorer
Edit
View
Favorites
Tools
Help
ess
http://128.82.7.208:9090/dtic/meta/LPS64487.xml
Go
Links
ogle
Search
2 blocked
Check
AutoLink
AutoFill
Discussions
Discussions not available on http://128.82.7.208:9090/
Done
Internet

```
<?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>
```

Benefits-Metadata Extraction

- Speed up the citation creation process
- Increase consistency in cataloging
- Improve quality in citation creation
- Reduce turn-around time in document processing
- Facilitate higher volume in document processing
- Improve job satisfaction for analysts by eliminating data entry job duties to focus more on intellectual content
- Improve retrieval of documents
- Integrate with EDOC (DTIC input processing system) with minimum effort

Current Status

- Development of the software for documents with RDP was completed and delivered to DTIC in June 06
- Testing of the software in the production environment has been completed and is in the process of integrating with DTIC input processing system
  - Benefits: Over 50% documents have RDP, and data entry to create metadata will be eliminated. Acquisition staff might not have to fill out the Web submission form.
- On-going:
  - Development of the software for documents that do not have RDP
  - Also switching software that automatically selects Template for each document type
- Future:
  - Develop “knowledge” base to improve the quality of the metadata output