IMPROVING COMPUTATIONAL USABILITY OF UNSTRUCTURED PILOT MEDICAL CERTIFICATION DATA

Sponsor: Federal Aviation Administration
Dept. No.: P233
Contract No.: 693KA8-22-C-00001
Project No.: 100976.10.102.1011.SD4
Outcome No: 4-5.B.1-6
PBWP Reference: AAM Strategic Roadmap 2022-2030

McLean, VA

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-23/22		

4. Title and Subtitle	5. Report Date
Improving Computational Usability of Unstructured Pilot Medical Certification Data	June 2023
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
C. Horowitz	Product 4-5.B.1-6

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
The MITRE Corporation 7515 Colshire Drive McLean, VA 22102	11. Contract or Grant No. 693KA8-22-C-00001

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	Technical Report
	14. Sponsoring Agency Code

15. Supplementary Notes

Author ORCID: 0000-0001-5270-4673

Technical report DOI: https://doi.org/10.21949/1528562

16. Abstract

Current Federal Aviation Administration's Office of Aerospace Medicine (AAM) operations include scanning paper documents received from various third-party medical providers to support individual pilot medical certification decision-related matters. These operations lack analytic tools, resulting in a time-intensive effort by subject matter experts to manually search through document sets to find relevant information. MITRE applied human language technologies to a sample set of AAM documents containing unstructured pilot medical certification data. MITRE demonstrated that third-party documents received and scanned by AAM could be automatically classified through a combination of computer vision and human language technologies; demonstrated that these documents can have their content extracted with sufficient accuracy to enable fundamental automation and human support tasks including summarization, search, and de-identification; identified aspects of the documents that present risk to future systems development; and produced a prototypical integrated document processing software pipeline capable of automatic ingest, classification, content extraction, and indexing to support basic search as an initial application.

17. Key Word	18. Distribution Statement
Aviation, Safety	Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	10	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

OVERVIEW

Problem Statement

The Federal Aviation Administration's (FAA's) Office of Aerospace Medicine (AAM) has enacted a data strategy for safety risk management and safety assurance, using data-driven approaches to improve operations and policy management. Current AAM operations include the mail receipt of paper documents sent by various third-party medical providers to support individual pilot medical certification decisions. While current operational levels were unavailable for this report, dating as far back as late 1999, hundreds of such mailings were received for processing each day. Electronic scans of these mailings are produced using AAM's Document Imaging Workflow System (DIWS) and stored in a simple case-based organization of image files. No capabilities for rapid retrieval or analysis of information contained in scanned images are available to medical examiners supporting safety management or safety assurance missions.

This limited system functionality results in a time-intensive effort by medical examiners to manually review document sets to find relevant information when forming individual certification decisions or responding to other requests or needs on a case-by-case basis. In many circumstances, when AAM's data strategy would call for conducting broader studies of medical certification data and decisions for policy analysis and other organizational functions, these important studies are simply infeasible and, therefore, not undertaken because of the resources required to manually find and analyze important information by expert staff.

Supporting a Data Strategy

To fully enact the AAM Data Strategy, the information in DIWS should be converted to useful data, enabling functions that support routine operations, such as information search, document summarization, case analysis, and monitoring and alerting. Such functions could speed the basic operations performed by staff, reducing overall costs and response times, allowing more time for critical information analysis and decision-making, and creating more consistent processes by which decisions are reached. Further, the various studies that could be conducted on a broader population basis, important to informing and validating policy making, if the information were available as computable data, could be conducted efficiently, iteratively, and regularly.

The collection of document scans retained in DIWS, both historical and contemporary, has been done without significant awareness of, or concern for, the machine readability of that information. As such, there is significant uncertainty within AAM as to whether that information can be successfully converted to useful data formats with sufficient quality and semantics to make it valuable for potential strategic capabilities, as discussed above.

Research Objective

AAM tasked the MITRE Corporation's Center for Advanced Aviation System Development (MITRE CAASD) to design and perform technical experiments applying human language technologies to AAM documents containing unstructured pilot medical certification data sampled from DIWS. These experiments' purpose, consistent with the AAM Data Strategy, was to evaluate whether available technologies can be used:

- to create automation or human support solutions that could improve mission and business performance for the organization responsible for the ingest and use of this data;
- to develop a representative system design for processing documents to extract the content into useful data; and,
- to identify technical challenges that may present significant risks to a future system's cost or performance.

Data Security

While this tasking anticipated the complete DIWS document collection being made available for analysis and experimentation, various information security and assurance constraints precluded sharing this document collection within the time constraints of the task. Therefore, AAM provided a sample of representative documents for use during the task, understanding that task results may not fully address the range of technical challenges to be encountered in future system development based on the idiosyncrasies of the general document collection not represented in the provided sample.

Additionally, it was decided that data should be processed in an isolated, offline environment to minimize the potential for sensitive data to be compromised in light of the potential delays in gaining access to and using FAA enterprise computing resources. Similarly, this was decided with the understanding that it may have limited the selection of tools that could rapidly be implemented in such an environment.

The presentation of results during the conduct of this task did not suggest any significant concerns with either of the risks calling into question the conclusions reached, given AAM's familiarity with the data and what information could be useful when made available as computable data.

APPROACH TO ANALYSIS AND EXPERIMENTATION

AAM provided MITRE with 205 PDF-formatted electronic files and 105 XPS-formatted electronic files to develop an experimentation plan and apply human language technologies. The documents were drawn from a subset of predefined document types used to broadly label each document ingested via DIWS.

These documents were manually reviewed and evaluated to identify key technical challenges with converting the information into useful, computable data and to consider the techniques and tools that might best be applied towards meeting those challenges. Each electronic file generally contained one or more full-page images of a scanned paper document produced by a third-party medical provider providing information pertaining to an airman's medical certification.

Initial Qualitative Evaluation

Electronic files were qualitatively and systematically evaluated based on the following:

- Image Quality – typical challenges with image quality in scanned documents include scanner artifacts such as scan shift, scale, and skew (i.e., warping in the photographic image of the original paper); margin noise and imager defects; greyscale thresholding; low resolution/blurring; inclusion of fax/document stamps; exaggerated use of color in the paper documents, and generally poor-quality paper. Image quality can affect basic recognition of text and graphic art (e.g., lines, curves, etc.) and alignment of components in a page layout, sometimes requiring pre- and post-processing techniques to be developed to maximize overall system performance.
- Layout and Content Consistency – typical challenges with the consistency of layout and content in document collections include differing versions of templates; form artifact interference (e.g., boilerplate vs. data, form lines, unfilled checkboxes, etc.); differing document production methods and technology resulting in alteration of fonts, text placement, etc., and mixed-mode composition (e.g., letters, reports, forms, etc.). Inconsistency within a document collection can limit the effectiveness of general-purpose tools to reliably extract information and convert it to computable data, sometimes requiring heavy customization or very advanced adaptable techniques (e.g., machine learning, etc.) to handle the diversity of the data more effectively.
- Information Quality – typical challenges with information quality include mixed-mode production (i.e., handwriting, machine print, annotations, graphics, and photos); mixed-mode content (i.e., narrative, ellipsis, data, and charts); domain-specific language (i.e., jargon) and informal standards for information requirements that lead to different levels of information quality and value. This can lead to the need for more information extraction and processing tools designed to handle different information styles, increasing system cost and complexity.

Document Classes

Based on the review of 310 documents within the sample set provided, MITRE characterized two coarse classes of documents:

Class A documents generally contain one or more of the following features:

- A relatively heavy presence of scanner artifacts.
- Most or all valuable information is conveyed through handwriting or hand annotation of form elements.
- Routine use of form templates (various versions and production methods; forms not always in order or complete).
- A mixture of boilerplate and airman information.
- Presence of fax/document stamps and margin notes.

Class B documents generally contain one or more of the following features:

- Typically letter or report-oriented multi-sectional document.
- Mixed-form content mostly featuring narrative text, elliptical texts, data in tabular form (with or without ruling lines) or key-value pairs, and charts and other graphical figures.

It was further determined that each DIWS document type represented in the sample strongly, though not exclusively, corresponded to a given class rather than being a mixture of the two.

Based on the above coarse classifications, a set of information extraction and processing requirements can be considered that will be useful to select and prioritize techniques and technologies as system development and application delivery are planned.

Table 1 presents the assigned course class for each document type represented in the DIWS sample set provided for this task.

Table 1. DIWS Document Type Classification

DIWS Document Type	Class	DIWS Document Type	Class
afib_status_report	A	narrative-neuro	B
diabetes_on_insulin_recert	A	narrative-pulmonary_lung	B
dmo_status_report	A	narrative-rheumatology	B
faa_seizure_questionnaire	A	narrative-vision_eye	A/B
gender_dysphoria_status_form	A	osa_status_report-initial	A
narrative-cancer	B	osa_status_report-recert	A
narrative-cardiac	B	pacemaker_status_summary	A/B

File Processing Workflow

Based on the document collection and the coarse classification described above, a conceptual system design was developed to define a file processing pipeline whose input is an arbitrary electronic file, ostensibly received via DIWS, and whose output is the computable data describing the content of the electronic file as needed to support a range of hypothetical user tasks. The most tangible and basic of these user tasks would be the ability to search for text across files and retrieve the relevant document page image(s) of interest in search results for efficient review.

The general approach to system prototyping was to develop a general-purpose system to minimize the number of assumptions expressed in the software that enables the overall system performance to increase and for those assumptions being expressed, to minimize their scope and impact. This approach ensures that the overall system remains robust when operating on diverse data, even though its performance may not reach the highest possible performance that an overly-specialized system could be made to achieve on a specific task for which it was precisely designed.

Figure 1 depicts the file processing pipeline developed to support iterative experimentation with different human language technologies, tools, and techniques to determine the general performance achievable with the given DIWS electronic file sample:

1. Identification of an arbitrary input file as a PDF or XPS for further processing or unknown file type to be discarded or otherwise set aside for human review. XPS files are converted to PDF for further processing through uniform tools in the remainder of the pipeline.
2. PDF content is extracted into computable form. Features extracted from each PDF page include text, the size and position of such text on the page (i.e., layout information),

graphic lines on the PDF page, and the scanned image for each document page from which these features were extracted.

3. Each PDF page is predicted to be a specific page of a known FAA form template from a catalog of relevant templates developed to support third-party medical providers participating in the medical certification process.
4. Automated triage is performed to verify that the prediction performed in the prior stage appears reliable (i.e., a PDF page indeed belongs to the predicted Class A document), is indicative of being incorrect or invalid (i.e., a PDF page belongs to a Class B document), or is sufficiently ambiguous such that the page should be queued for a human reviewer to make a manual determination.
5. For each PDF page classified as Class B,
   a. Perform text analysis to automatically determine the DIWS document type that should be assigned to this page;
   b. Perform layout analysis to identify distinct document regions (e.g., section headers, paragraphs, data tables) for which to apply different extraction techniques; and,
   c. Perform automated annotation of text for basic linguistic analyses and entity recognition (e.g., persons, places, dates, clinical domain terminology and concepts).

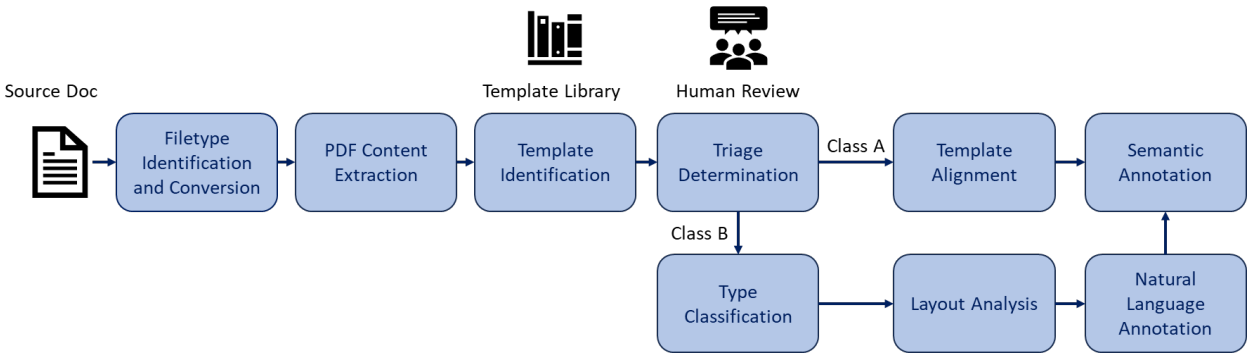

Figure 1. File Processing

6. For each PDF page classified as Class A, use alignment techniques to identify template zones within the page that are likely to contain information entered by the provider such that those zones, separate from template boilerplate, may be analyzed and information be extracted.
7. Apply semantic annotation of extracted text to identify higher-level concepts such as events and topics to enable further indexing and analyses.

The output of this pipeline is textual, graphic, and semantic data which could be indexed to enable search functions to be delivered. Search functions may include basic text search, searching for the presence of features in a document (e.g., topics, form annotations), and searching for semantic data, including entities (e.g., people, places, medicines) and events (e.g., diagnoses, dosage changes, tests).

Performance Evaluation

Following the initial prototyping for each stage, an evaluation was performed to estimate the stage's performance using the sample DIWS electronic file set. The typical metric for the

performance of these capabilities, known as an *f-score*, is a combination of precision (i.e., the quality of the system's relevant outputs) and recall (i.e., the quantity of the system's relevant outputs). The evaluation of each stage's initial prototype found that initial prototyping fell in the 80% to 100% range. In many cases, iterative development showed that f-scores could be marginally improved with each iteration by prioritizing and addressing the residual nuances with specialized or custom techniques.

Analytic system performance can generally be improved through further development, and in some cases, by introducing newer technologies and techniques over time. In the case of this prototyping effort, the ability for contemporary tools and techniques to reach this level of performance in the first attempt to process this data should be viewed as confirmation that the electronic files contained in DIWS can be made useful without extensive new technology creation. The shortcomings in system performance will generally yield the consequence of textual data containing errors based on incorrect recognition (i.e., incorrect text data) or failed recognition (i.e., missing text data).

Provided the applications delivered to users, and the processes that dictate their use, are sufficiently tolerant to the potential quality problems, this should not be viewed as a barrier to using the electronic file collection. Tolerance can be created by techniques such as providing users with hybrid tools that enable them to rapidly view the original document images in line with searching system data or producing population studies to confirm the quality of the system data.

Technical Challenges

During the performance of this task, the following technical challenges were identified as significant enough to warrant specialized consideration:

- As noted above, handwriting is a regular feature within the Class A document set. In most cases, it is the primary or sole means by which the provider conveys information. Handwriting recognition technology is becoming more rapidly available, both in terms of the availability of basic tools and via online commercial services. However, this technology is not yet to a degree where it can successfully process arbitrary handwriting from a diverse population of authors. Additionally, due to the nature of the documents collected in DIWS, some pre-processing would be required to make the handwriting images of sufficient quality for recognition technologies to be successful. During experimentation, it was demonstrated that handwriting in a given region or zone of a document page might be detectable in many cases, and that could prove to be useful to users in speeding the retrieval and analysis of information from documents. However, correctly recognizing that writing will remain a challenge that requires focused investment to overcome gradually.
- As also noted for Class A documents, the general image quality for these materials is generally lesser than for Class B documents. This may result from the fact that Class A documents are often prepared mainly by hand and, as such, undergo a greater degree of handling during their production, including being scanned by the provider as part of their record-keeping activities. This may lead to the degradation of the original paper and the possibility of AAM receiving a scanned copy of the document, which might feature scanner artifacts from local provider equipment. This phenomenon is likely to continue so long as the processes and resources for documents being received by DIWS encourage the production of documents by hand. Investing in stakeholder engagement, publishing revised

resources and guidelines, and enhancing electronic processes may be beneficial in diminishing the amount of low-quality document imagery that is ingested. There is also the potential that DIWS is contributing to the quality problems seen in Class A documents, and, therefore, some systems analysis may be required to determine if some system modernization may affect data quality, as well.

RECOMMENDATIONS

MITRE recommends that AAM proceed with development to quickly create value for operational users and simultaneously develop strategic capabilities requiring more extensive customization to achieve AAM's mission and business objectives. While the identified technical challenges will be a source of potential investment as system development plans are formed, most of the technical activity can be focused on delivering value to users via new functions that will quickly improve their performance. Additionally, the ability of the electronic file collection to support population-wide studies that can contribute towards higher degrees of safety management and assurance represents a distinct value that can be realized in parallel with creating those basic operational capabilities.

The following key activities are suggested to enact this recommendation:

- Apply the AAM Data Strategy to the ongoing acquisition of medical certification-supporting documentation by engaging stakeholders, particularly third-party medical providers responsible for the production and submission of such documentation, in considering how to improve that information from the standpoint of ensuring its machine-readability while simplifying processes for production and submission consistent with contemporary methods in medical provider facilities, where possible.
- Continue developing an integrated document processing pipeline to extract information into forms usable in computing applications, including classification, search, summarization, tabulation, and complex analysis. Specific areas of development that could increase the value of work done in the prototype include additional pre- and post-information extraction processing techniques; inclusion of task-specific assumptions, domain tuning, and subject matter expertise in extraction and analysis task implementation; and formulating specific application requirements to help prioritize and define critical objectives for system operation and performance. Based on the course classifications described in this report, it is recommended that document types classified as Class B documents be prioritized for application delivery projects. The general image quality found in these document types will allow delivery projects to focus on user functionality and better measure application performance and effectiveness. The density and richness of information in these document types may also lead to significant operational performance improvements as information retrieval and analysis tools are made available to process these documents. It is further recommended that within Class A documents, the documents typed as *afib_status_report* within DIWS be utilized for risk reduction projects, such as the development of information extraction tools and potential work on handwriting recognition. This document type is recommended because, within the set of Class B document types, it contained the best balance of challenges with consistency and quality such that techniques and technology can be developed without being unnecessarily

complicated by too many coexisting technical challenges, particularly for early risk reduction efforts to show progress and allow AAM to better calibrate for future work.

- Identify or establish a secure and scalable computing environment capable of hosting the complete document collection and delivering sufficient computing resources to perform rapid, iterative development of the integrated document processing pipeline for experimental, development, and operational needs.
- Manage identified technical risks through selected investment in further technology experimentation to improve and extend information extraction, revision of organizational processes associated with medical certificate data acquisition to reduce future technical burden through improvements to data quality, and rapid development of applications to validate user needs and organizational benefits.