International Civil Aviation Organization
CAR/SAM Regional Planning and Implementation Group (GREPECAS)

INFORMATION PAPER

GREPECAS/20 — IP/16
27/10/22

Twentieth Meeting of the CAR/SAM Regional Planning and Implementation Group (GREPECAS/20)
Salvador, Brazil, 16 – 18 November 2022

Agenda Item 7: Other Business

INNOVATIVE TECHNOLOGY USE IN THE EXTRACTION OF FLIGHT CONSTRAINTS RECORDED IN LETTERS OF AGREEMENT (LOA)

(Presented by the United States)

EXECUTIVE SUMMARY

Foundational to ICAO's vision of an integrated, harmonized and globally interoperable Air Traffic Management (ATM) System, as described in the Global Air Traffic Management Operational Concept (Doc 9854), is the availability of ATM System information supporting collaboration and automation. Operational constraints are a subset of ATM information of particular importance to Airspace Users (AU) due to the impact to their planning and responses to unforeseen circumstances. The advancement of Machine Learning and Data Analytics techniques have led the United States Federal Aviation Administration (FAA) to begin exploring uses of these technique to extract constraints information. This paper details the different Natural Language Understanding (NLU) techniques and relevant tools the FAA has used to to extract the constraint information from the text in Letters Of Agreement (LOA) documents.

Strategic Objectives:	• Air Navigation Capacity and Efficiency
References:	• Doc 9854, *Global Air Traffic Management Operational Concept* • Doc 9965, *Manual on Flight and Flow – Information for a Collaborative Environment* (FF-ICE) • Doc 9971, *Manual on Collaborative Air Traffic Flow Management* (ATFM)

1. Introduction

1.1 Foundational to ICAO's vision of an integrated, harmonized and globally interoperable ATM System as described in the Global Air Traffic Management Operational Concept (Doc 9854) is the availability of ATM System information supporting collaboration and automation. Operational constraints are a subset of ATM information of particular importance to AUs due to the impact to their planning and responses to unforeseen circumstances.

1.2 Letters of Procedure/Letters of Agreement (LOPs/LOAs) contain procedures that are supplementary to ICAO Standards and Recommended Practices in Annexes 2 and 11, the Procedures for Air Navigation Services – *Air Traffic Management* (Doc 4444) and Regional Supplementary Procedures (Doc 7030). They describe conditions, responsibilities, and other aspects of agreement across Air Traffic Services (ATS) units regarding service provision expectations. The specific agreements, in many cases, result in operating practices that constrain how an individual flight may pass from one ATS unit's jurisdiction to another.

1.3 Today, AUs are generally not aware of some operating conditions and expectations (e.g., Instrument Flight Procedure Crossing Constraints) contained in LOAs which might constrain flight trajectories. In the case of the FAA, cross-ATS unit agreements are documented in over 20,000 LOA Portable Document Format (PDF) documents. The documents include both text as well as graphics and scanned pictures that are not publicly available in a consistent or digital (machine useable) manner. As a result, the impact of applicable LOAs on flight planning or active flights is only understood by the AUs after a flight plan is filed or once the aircraft is airborne.

1.4 In many cases, AUs "intuit" the LOA parameters from years of experience. Should an LOA be amended, AUs are forced to identify the new, unanticipated changes to their flights to understand the updated constraints contained in the LOA, then reinterpret the parameters they use for flight planning or airborne requested changes. This reactive method is inefficient, affects flight planning, causes reroutes, and increases workload and complexity for both the Air Navigation Service Provider (ANSP) and AU. An alternative to AUs trying to discern LOA changes and associated impacts through trial and error, the ANSP could provide applicable constraint data contained in LOAs in a digital, machine useable format for ingestion by AU automation support to better inform their planning as well as their participation in Collaborative Decision Making processes.

1.5 The challenge in extracting information from these documents is interpreting the information from their semi-structured natural language form. In recent years, the advancement of information extraction in NLU techniques and relevant tools have become increasingly available to tackle this challenge. The FAA has undertaken an effort with the United States National Aeronautics and Space Administration (NASA) to extract this flight constraint information from the text in LOA documents using the modern NLU techniques. The constraint information resulting from the extraction/identification process serves as input patterns for inclusion in a current, standard schema and exchange model (e.g., Aeronautical Information Exchange Model (AIXM)). The initial focus is on constraint information contained in LOAs for FAA Air Route Traffic Control Centers (ARTCCs)/ICAO-Area Control Centres. This paper presents the results and progress made thus far.

2. Discussion

2.1 Data Selection

2.1.1 To extract information most impactful to AUs, it was decided to focus on isolating constraints with direct impact to the aircraft and flight operations (i.e., flight constraints). This requires differentiating agreements with direct flight impacts from other aspects of cross-ATS unit agreements, such as required coordination or the manner in which the coordination is accomplished. Relying on the general format of the LOA PDF documents, the "Procedure" section was identified as most likely to contain flight constraints. Therefore, the process was designed to extract information from the individual lines within each "Procedure" section.

2.2 Overview of Process

2.2.1 NASA designed a pipeline process to include steps utilizing Machine Learning (ML) techniques (i.e., Name Entity Recognition (NER) and clustering) and traditional scripting to extract the flight constraints from the procedure lines of the LOA PDFs, as shown in Figure 1.

Figure 1: LOA Constraint Extraction Pipeline

2.2.2 Constraint Identification and Extraction

2.2.2.1 To achieve the goal of extracting generalized flight constraints, the following steps were proposed and implemented. First, the text and key sections of the document were extracted using PDF2Text. Second, the traditional dictionary gazetteers[1] and syntactic lexical pattern matching were used to extract named entities. ML was used to assist in flight constraint identification by performing pattern recognition. Lastly, pattern templates were generated from the data samples to be fit and extracted from the text.

2.2.2.2 Throughout these steps, Subject Matter Experts (SMEs) conducted continuous manual validation of the intermediate results to refine the model development process.

[1] A gazetteer consists of a set of lists containing names of entities such as cities, airports, aircrafts, organizations etc. Lookup methods with the gazetteer are used to recognize entities in the text.

2.3 Discussion of Techniques

2.3.1 Named Entity Recognition (NER)

2.3.1.1 The LOA PDFs were inputted into the Application Programming Interface and JavaScript Object Notation files with the text for each line of the LOA that were returned. A custom parser then extracted each line from the "Procedure" sections and performed formatting to link it with the relevant headers and sub-headers in the LOA. NER, a form of Natural Language Processing (NLP), was used to support the labelling process. The set of entities to be labelled was defined to capture aviation domain-specific information that would be present in flight constraints (e.g., aerodrome and ARTCC names, block altitudes, fixes, etc.).

2.3.1.2 A combination of dictionary lookup with pattern matching was used to label the entities present in each line. These techniques were chosen due to the standardization of domain-specific language used in LOAs which yielded good results. NER can be used to extract patterns within text given a broad enough data set that has inherent patterns.

2.3.1.3 A trained NER model identifies keywords within a document and higlights the the keywords throughout the document. For example, LAX would be tagged as an airport by the model in the sentence, "The aircraft departed LAX." Labelling is an intensive process. It is important that it is consistently done so that the algorithm can learn new keywords.

2.3.1.4 Having consistent labels and the appropriate number of labels (or tokens) allows the NLP tools to better understand the structure between the words. Further, NLP tools can treat words or tokens with the same properties as similar even if they are different words in the original documents (e.g., jets and turboprops would both be labelled as aircraft). Certain labels (e.g., airspace and sector) were not used in this analysis and the potential to add these labels may change the results.

2.3.2 Clustering

2.3.2.1 To discover the unknown flight constraints contained in the LOAs, iterative unsupervised ML (i.e., clustering) was used to group flight constraint patterns from the LOA texts. This grouped similar lines together, allowing FAA SMEs to identify and label patterns that represent common constraints present in the data. The remaining unlabelled lines were then clustered again, and the process was repeated.

2.3.2.2 The clustering method discovered that sentence structure and the constraints contained in the LOAs were more unique than previously thought. Given the diversity discovered in the clustering, NLP was used to extract the subject – verb – object of each line to begin to extract the constraints.

2.3.2.3 Term Frequency Inverse Document Frequency (TF-IDF) is a simple metric that calculates how many times a word appears in a document, negatively weighted by how often it appears in all relevant documents. Using TF-IDF, LOAs can be represented by a list of values for every word. These vectors can be fed into clustering algorithms and then plotted on a graph to look for patterns within the data. The clusters were analyzed to determine if patterns were discernible from the cluster. However, the patterns that were generated from the clusters ended up being very generic and it became difficult to discern the differences between the patterns, as well as the details within the constraints.

2.3.3 NLP with SpaCy

2.3.3.1 SpaCy is an open source library that includes many NLP techniques. SpaCy was used to find the verbs of each line and begin to map the relationships of the words within the lines. With identified structure of the words, the information should be able to be moved into more flexible descriptions of the constraints. With a map of the word relationship and the associated tokens, it is believed that structures will be able to be mapped to elements within the AIXM schema. This final mapping is future work after the constraint extraction is complete.

2.4 Results and Lessons Learned

2.4.1 In the document preparation steps, the “Procedure” sections were successfully extracted from the LOA documents. From these “Procedure” sections, 28,911 formatted lines were extracted. This created a very well-formed dataset from which to label entities and extract flight constraints.

2.4.2 The entity labelling likewise performed very well. The rigid nature of gazetteers and pattern matching did miss some entities and as it was not able to differentiate between two different entities that are referred to by the same name. However, due to the structured nature of the LOA documents, very little was missed or incorrectly labelled.

2.4.3 Rule-based approaches seemed to thrive when performing NER. It is believed this is due to the formal writing standards provided by the FAA and the consistency of key terms like facility names and flight parameters (e.g., altitudes and speed). With a lack of initially labelled data, it also provided a reasonable baseline to work off when refining the final models.

2.4.4 Although key terms are typically written very consistently, the same is not true for overall document structure. Each ARTCC had a unique style of writing documents, and there is even variety found within documents posted by each ARTCC. This diversity in language and structure contributes to noise and makes ML tasks difficult, but also gives motivation to discover ways of standardizing LOAs into a digital format.

2.4.5 The diversity in the material contained in a LOA currently makes NLP the best fit to extract information from the labelled lines. The identified patterns were too specific to have enough matching lines to be a true “pattern.”

2.4.6 The extraction of constraint data from the LOAs is very challenging, and the initial efforts are promising within their constrained context. The techniques discussed in this paper offer an opportunity to support the creation of digital information about the procedures and processes of ASPs that currently impose opaque flight constraints on AUs.

3. Conclusion

3.1 The Meeting is invited to note the information provided in this paper.

— END —