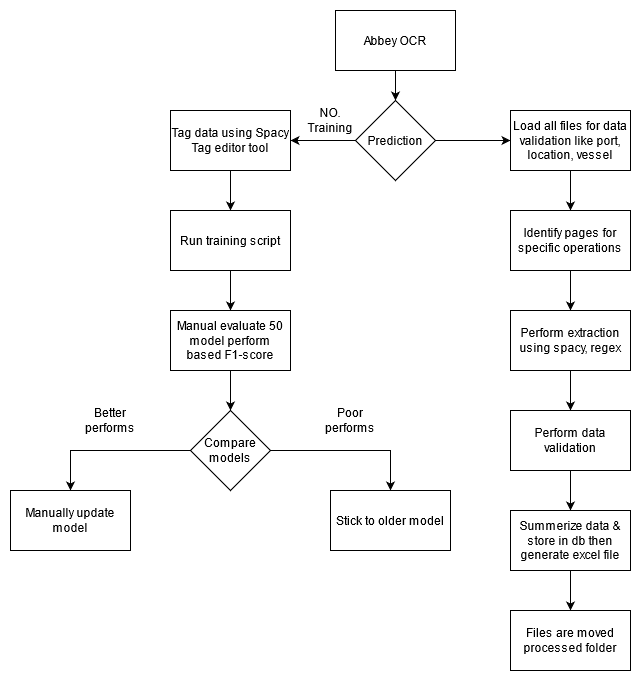
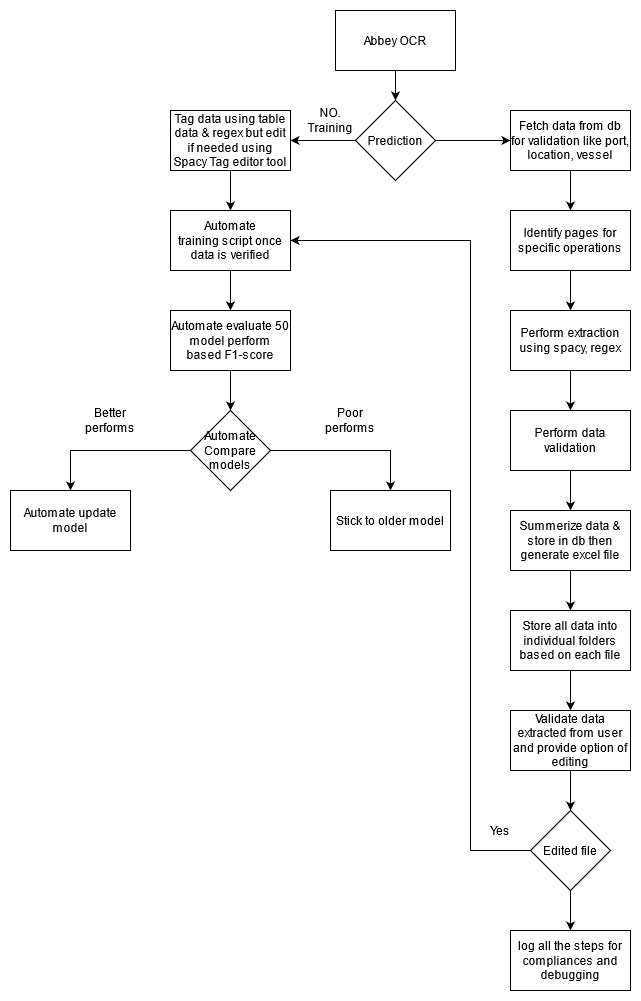
**Current flow**

****

1. Once the file placed into the target folder UI path will consume it for processing through Abbey OCR reader for generating text files.
2. Tag data using spacy tag editor then manually save that file as a text file.
3. Run training script manually for 50 different models then evaluate each model performs based on f1-score of each entity.
4. To perform prediction load all the necessary files of geolocation, vessel, port entities.
5. Load bag of words file for port and city name for removing confliction found in documents.
6. Generate folder for a specific date then move files into them then perform extraction like the seller, buyer, product, etc based on the page type.
7. If page type is not applied then extract port,other location, and person name-related information.
8. Product name if the product name is not present inside the application page.
9. Fetch country name from NER model then compare with geolocation file for validation.
10. Combine all port information into a single data frame like exporter, buyer, and other locations.
11. Email extraction using regex.
12. After all, data have been extracted create 2 files complete and summarized excel file before summarization combine allorganization names.
13. All this data is then inserted into DB.

**Proposed flow 1**

****

1. Once the file placed into the target folder UI path will consume it for processing through Abbey OCR reader for generating text files.
2. Tag data using spacy tag editor then manually save that file as a text file so instead of doing this, we can use our geolocation, vessel, port file and salutation/regex for finding name then tag data automatically just cross-check from users to save time.
3. Once verification completed then move into a specific folder to start the training process automatically.
4. Once training has started for 50 different models then evaluate each model performs based on f1-score for each entity then decide the best model by automatically comparing with the previous model performs. All model performs should be stored inside DB as part of logs.
5. To perform prediction fetch data from tables like geolocation, vessel, port, etc entities as per requirement.
6. Load bag of words file for port and city name for removing confliction found in documents.
7. Proposed folder structure like a unique id-based folder for each pdf file then inside them folder for individual steps finally output excel file.(reflect unique id on screen)
8. Fetch data like port, other location, and person name with other related information if page type is an application.
9. Fetch data like product name if page type is not an application.
10. Fetch data like a country name from NER model then compare with geolocation file for validation.
11. Combine all port information into a single data frame like exporter, buyer, and other locations.
12. Email extraction using regex.
13. Once all information is collected from text files then system generates 2 excel files one contents information with repeations in them and another contents summerized form of it.
14. All this data is then inserted into db.
15. Validate extracted data by providing UI to end-user for cross-checking and option of updating extracted information with maker checker. Updated information is sent to further processing and will use to improve performes of the system.
16. Maintain logs for each step of the processes to identify completed stages and occurrences of a failure.

**Advantages**

1. Automatic data tagging using available information leads to reduction in tagging time.
2. Manual training process will be removed so time saving for report generation and evaluation of the model and it’s updation.
3. Moving from file base loading to DB base data loading leads to secure information.
4. Multiple methods of extracting data helps us to hand failure of other method.
5. Editing option to improve result by manual intervention that’s better data tag and repeation of training cycle.
6. Maintaining logs of each step of the process to keep track of processes.

**Disadvantages**

1. Manual cross checking of data properly tagged or not and in case of failed to identify entities then tag them manually.
2. Keep up to date the DB tables of keywords to perform automation in tagging.

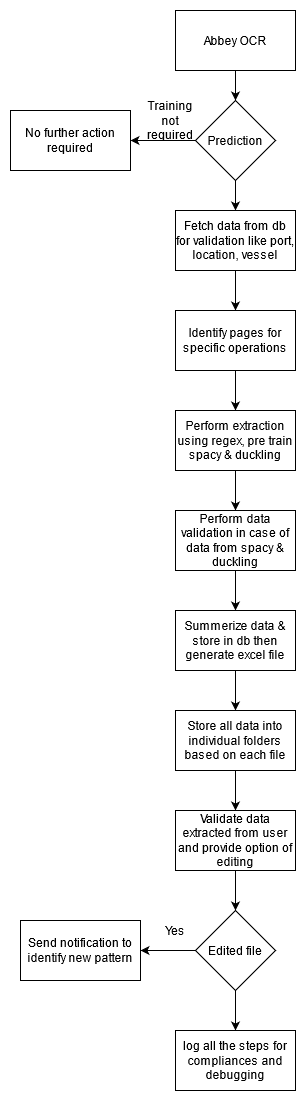
**Benefits for end user**

1. Option of editing incorrect extract value.
2. Complete flow can be handle by single pipeline.

**Benefits for IT**

1. No more requirement of complete manual tagging.
2. Multiple methods to extract same value so less chances of failure.
3. Easy to debug issues due to restructuring codedbase.
4. No manual trigging for training or retraining it will be completely automate using tools like AirFlow.

**Proposed flow 2**

****

1. Once the file placed into the target folder UI path will consume it for processing through Abbey OCR reader for generating text files.
2. Completely remove the process of tagging but instead, we can use our geolocation, vessel, port tables, and salutation using db regex method for finding entities.
3. For general entities like geolocation, date, the amount we can still use spacy + duckling libraries with their pre-trained models.
4. To perform prediction fetch data from tables like geolocation, vessel, port, etc entities as per requirement.
5. Load bag of words file for port and city name for removing confliction found in documents.
6. Proposed folder structure like a unique id-based folder for each pdf file then inside them folder for individual steps finally output excel file.
7. Fetch data like port, other location, and person name with other related information if page type is an application.
8. Fetch data like product name if page type is not an application.
9. Fetch data like a country name from the NER model then compare with geolocation file for validation.
10. Combine all port information into a single data frame like exporter, buyer, and other locations.
11. Email extraction using regex.
12. Once all information is collected from text files then system generates 2 excel files one contents information with repeations in them and another contents summerized form of it.
13. All this data is then inserted into db.
14. Validate extracted data by providing UI to end-user for cross-checking and option of updating extracted information. Updated information is sent to further processing and will use to improve performs of the system.
15. Maintain logs for each step of the processes to identify completed stages and occurrences of a failure.

**Advantages**

1. No need of tagging data only general purpose pre trained model will be used of libraries like spacy, duckling.
2. Moving from file base loading to DB base data loading leads to secure information.
3. Multiple methods of extracting data helps us to hand failure of other method.
4. Editing option to improve result by manual intervention that’s better data tag and repeation of training cycle.
5. Maintaining logs of each step of the process to keep track of processes.

**Disadvantages**

1. Keep up to date the DB tables of keywords to perform automation in tagging.

**Benefits for end user**

1. Option of editing incorrect extract value.
2. Complete flow can be handle by single pipeline.

**Benefits for IT**

1. No more requirement of complete manual tagging.
2. Multiple methods to extract same value so less chances of failure.
3. Easy to debug issues due to restructuring codedbase.
4. No manual trigging for training or retraining it will be completely automate using tools like AirFlow.

**Data Validation Screen**