Project specifications:
[AI Reader CAM]

Contents

1. Introduction

1.1 Name of the Project

AI Reader CAM

1.2 Contracting Authority

Commissariat aux affaires maritimes (CAM)

1.3 Type of Solution Required

Development of IT solution, and an adaptable prototype.
The solution should be a proof of concept (POC) which will later be adapted for a productive usage. This proof should show that all the requireed expectations on the input and output data can be realized and that the expectations on the core functionality are met. The functionality expected to be present in the POC will be highlighted in the document.

1.4 Responsible Project Team

Name	Organization	Service	Role
Frederic Hellenbrand	Commissariat aux affaires maritimes	IT	Responsable IT
Noe Wolter	Commissariat aux affaires maritimes	GdM	Responsable GdM

2. Execution of the Contract

2.1 Technical and Specific Description of the Required Solution

2.1.1 Context and Objectives of the Required Solution

The Commissariat aux affaires maritimes has started a new round of modernization and digitalization.

This project is long term and includes:

- A rewrite of the operational application suite GESTCAM (from 1998) which manages the processes and data of the main services through GESTNAV, GESTMARIN and GESTDIR.
- Digitalization of documents and certificates issued by the CAM.
- Streamlining and automatisation of the processes , with an emphasis on data acquisition of information contained in electronic documents, but not fully dematerialized, as we know that full digital payloads will not be provided by all actors in the maritime field for several years to come.

The majority of these efforts depend on the progress of the rewrite of the management application suite. But on the topic of data acquisition there is this project, subject of this call for solution: a tool that allows predetermined data to be extracted from PDF (text or images) or JPEG documents without human intervention. The goal, in a first phase, is to improve the efficiency of the data entry processes for the "Seafarers" service. This involves reducing the number of data points submitted by our clients, decreasing the number of errors occurring during this step, minimizing the need for verification and correction by CAM agents, and enabling the automatic injection of data into the relevant systems.

2.1.2 The Required Solution

We are looking for an application which will extract data from official documents that our clients send us and facilitates the verification and input of this extracted information into a subsequent system.

The application must be trainable to identify the targeted texts in form of label/value pairs, in all types of documents and their variants. Ideally, it should be capable of self-training to identify label/value pairs in cases where it has already been trained to find values in multiple variations of the same type of document. We have knowledge that some solutions, do not need a manual definition of each search zone for each single document variant.

Through data validation by our agents, the system will improve over time as inconsistencies are reported. In some cases, the agent may need to provide the exact search rectangle, but this should be the exception rather than the rule.

Since the program will process documents containing personal data, GDPR compliance must be considered. The application must not use third-party public cloud services to process documents and data. The CTIE offers VMs with GPU support running Linux for any application using Machine Learning and AI.

In the first phase, two types of documents must be readable and their data extractable: VISAs (STCW certificates and endorsements) and seafarer booklets (passports). A sample of the documents and the expected output is available in the annex documentation.

2.1.3 Target audience

The solution will be used by the CAM agents. No prior knowledge in programming is assumed for the users of this solution.

There should be an installation and configuration manual available for IT Admins to deploy the solution to any suitable server instance and add new types of documents. The training of the variants of a new

types of documents should be feasible by qualified personnel, who are familiar with machine learning. Solutions which do not oblige the CAM to have further software development for extending the document types, will be preferred.

2.1.4 Work and Work Environment

1. Initial development

The development of the solutions will first have to be done on the provider's premises with anonymized datasets. Once a deployable and beta ready solution is available, the solution will be installed on the CTIE's servers, known as DEV, as to ensure that the solution can be run on CTIE's servers until the CAM's IT department and GdM service are satisified with the prototype. The environment will be provided on OS level as required by the provider.

2. Model training and user acceptance tests

   A UAT test system will be setup, which will contain live data, but will only be accessible through CAM's premises. This system will be trained to the specific types and their variants with production data. No live data will leave the CAM or CTIE's systems. The UAT phase will require the provider to come on site to proceed to the installation configuration and assist in the training of the platform.

3. Integration into actual business processes

As required in this POC, the module will be integrated in the business processes by giving the user's the possibility to drop the documents to be analyzed in dedicated shared folders, where the application will pick up the documents and extract the data. The data will be available in the GUI front-end, where the user's can validate the data and eventually modify it to 'educate' the model. From there the user will copy/paste the data into the target applications.

4. Integration in the future GESTCAM application

The POC should bring proof that other input and output possibilities are possible, either by mocking such the future endpoints or deliver convincing code, which will be evaluated by CAM IT or an external party. CAM IT will integrate the module in a later stage, out of scope of this project, with the help of the provider of the GESTCAM solution and the CTIE.

CAM IT will provide the VMs with the required OS Systems. The application itself will whether be installed by the provider or will be installed by IT CAM, if a complete installation and configuration manual is

available. A solution running under Linux is preferred, but system admin competency is available for linux and windows servers.

The CTIE provides specialized GPU server VM in its service catalog.

Management and milestone meetings can be held online via Teams or at the CAM's premises. Model training, update and expert meetings will be held on site preferably.

It must be mentioned that the Ministry for Digitalization will also be involved in giving feedback throughout the project.

2.2 Data and Technology Specifications

2.2.1 Data Specifications:

The data are values to be extracted from two document types for this phase:

- Certificates and endorsements
- Passports

The CAM will provide as many variants as possible for each of the document types. Those variants differ slightly in design depending which institution or country is the emitter. In all cases a common set of fields can be found on each document. It is those pairs of label / value which will have to be detected in the documents. Most pairs appear a single time, but some appear in tables and thus can be present 1-n times.

2.2.2 Technology requirements (internal and external)

The program needs an AI or ML driven engine, capable of "reading" documents and extracting key data from it.

The expected outcome of the project is an autonomous application, ideally with a local client or a web front-end and a server backend. The incoming documents would be made available through either:

- a drop in shared folder -> required for the POC
- an upload via the UI interface, ideally drag and drop. -> required for the POC
- an upload per sftp to a shared folder -> proof of feasibility
- injection through an API -> proof of feasibility

The data extracted from the documents should then either:

- be presented to the user in an orderly and formatted manner on-screen, allowing it to be easily copied/pasted. A log of the extraction should be saved on disk. -> required for the POC

- or directly injected into the operational application via SQL or through a dedicated API. The latter option is not optional, but will only be required and implementable when the new management application suite is available. -> proof of feasibility

The program will run on a server in the Govcloud provided by CTIE. The input data will come from Guichet Unique under PDF or JPEG format. The output data will be multiple fields returned as label / values and in tables for the data which has 1-n instances.

The authentication to the application's frontend should use the IAM/TAM authentication system provided by CTIE. Only personal with a IAM account should be managed.

The program will need to comply with RGPD requirements.

2.3 Expected Deliverables

The deliverables are expected:

- An AI or ML powered engine, capable of analyzing the documents and extracting the data required running on a server.
- After an initial training on a document type and a subset of its variants, the solution should be capable of self-training for other variants of the same document type.
- A client interface allowing easy drag and drop to analyze the documents (no saving required).
- The possibility to configure the solution as to pick the documents from various sources mentioned under 2.2.2. Technology requirements.
- A client interface to verify and confirm the output and allowing for corrections. Corrections should be part of the learning process for the AI.
- The output must come in a format allowing for importation and integration with other systems, as mentioned under 2.2.2. Technology requirements.
- The solution must come with a configuration and installation manual fo IT admins.
- Generic enough for future modifications, adaptations and application to other documents.
- Generic enough to be deployed on any server instance by following a configuration manual. Parameters should be configurable through configuration files by qualified IT admins at least.
- An overall metric for success is that the process of analysis, data extraction and importation of data into another system needs to be faster than doing the same work manually. It should require as little manual interventions as possible.

2.4 Profile of the economic operator

Profile of the economic operator:

Requirements:

- Company active in the field of software development with a proven record of successfully develop, deploy and maintain applications.

Criteria:

- Provide a track record of similar projects completed using text extraction and ML, with short explanation

Profile of proposed project team:

Requirements:

- Language requirements: French or English or German. All code comments and guides to be preferably written in English.
- Development to occur outside of CAM.
- VPN access to DEV and TEST servers can be provided.

2.5 Duration of the project

The timeframe is 6 months from kick-off to deliverables.

2.6 Rights of inspection

In order to facilitate both the transfer of knowledge to the contracting authority and quality control, the contracting authority authorises itself a right of inspection over all deliverables, methods and practices used by the economic operator during the production of the results inherent in these project specifications. The proposed solution and approach must take into account the fact that at the end of the contract the contracting authority wishes to be able to guarantee the maintenance of the project itself. During the project implementation phase, the economic operator grants a right of inspection to the contracting authority in order to allow it to monitor the quality and control the progress of the work.

2.7 Security, confidentiality and protection of personal data

The security and confidentiality aspects are taken seriously by the contracting authority. A non-disclosure agreement, a personal data processing agreement and a non-compete clause need to be signed by any person working on this project as well as by the economic operator. Technologies implying a transfer of the contracting authority's data to private cloud services are not allowed in the project. Government cloud

service (govCloud) will be used. The source code and all project deliverables will be made available to the contracting authority.

In the event of the processing of personal data on behalf of the contracting authority, the economic operator undertakes to conclude with the contracting authority a data processing agreement compliant with the provisions of Article 28 of GDPR. In addition, the contracting company must provide evidence that it provides sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing of personal data covered by this project will meet GDPR requirements, that data-protection by design is respected and that the protection of the rights of the data subjects in accordance with the GDPR are ensured.

Throughout the whole project and beyond, all employees of the economic operator involved in the project shall agree to maintain the confidentiality of all information and data in relation to the project and shall not use, disclose, transfer, or make it accessible to anyone other than authorised employees.

2.8 Property of the Solution

The results inherent in these specifications, algorithm, solution, results as well as the data generated are protected by the relevant intellectual property and copyright laws and will remain the exclusive property of the Luxembourg State ("Etat du Grand-Duché de Luxembourg") for the duration of the project and beyond.

The contracting authority grants no license or authorisation regarding the intellectual property rights, which it holds in respect of the project. The economic operator does not have, without the prior written consent of the contracting authority, the right to use the results to market them.

2.9 Communication about the Project

The economic operator does not have, without the prior written consent of the contracting authority, the right to advertise the services provided within the framework of this project. In the event of communication about the project, the economic operator shall mention in addition to the contracting party, also the Ministry for Digitalisation and the Tech-in-Gov Call for Projects.

3. Award Procedure

3.1 Applicable laws and regulations

Unless otherwise provided herein, the following laws and regulations, as amended, shall apply:

- la loi modifiée du 8 avril 2018 sur les marchés publics (L MP) ;
- le règlement grand-ducal d'exécution du 8 avril 2018 de la loi modifiée du 8 avril 2018 sur les marchés publics (RGD MP) ;
- le règlement grand-ducal modifié du 27 août 2013 relatif à l'utilisation des moyens électroniques dans les procédures des marchés publics et les procédures d'attribution de contrats de concession ;

- la loi du 23 juillet 1991 ayant pour objet de réglementer la sous-traitance (L ST) ;
- la loi du 10 novembre 2010 instituant les recours en matière de marchés publics (L RMP) ;
- la loi modifiée du 8 juin 1999 sur le budget, la comptabilité et la trésorerie de l'Etat (L COMP) ;
- les prescriptions du Code Civil (C CIV) ;
- le règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (GDPR) ;
- la loi du 1er août 2018 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel en matière pénale ainsi qu'en matière de sécurité ;
- la loi modifiée du 31 juillet 2006 portant introduction d'un code du travail (C TRAV) ;

3.2 Acceptance of the project specifications

By submitting an offer, the economic operator acknowledges having gathered all the information necessary to establish a valid offer, i.e. that he is aware of the difficulties and particularities of the to be solutions to be provided and takes them into consideration when preparing an offer.

Any modification to the specifications by the economic operator is considered null and void (RGD MP Art. 60 (1)).

3.3 Content of the Offer

- Brief description of the company and track record in the development of text extraction and machine learning applications .
- A description of the project management, analysis and development processes used and applied to this project.
- A detailed description of the architecture of the solution : Modules, frameworks, specific libraries and software languages used.
- A detailed list of licenses to obtain if those are needed. Ideally a breakdown of the license costs if those are bought through the provider and the cost forwarded to the client.
- A breakdown of deliverables

3.4 Price Regime

The contract falls under the fixed global price regime.

3.5 Submitting an Offer

The offer must be sent via email to cam@cam.etat.lu.