DIGITAL_TAX Service - Task Description

Problem: The National Revenue Administration (NRA), despite the development of public services supplied by electronic means of comunication, has to process tax information submitted by NRA customers in paper form. **In the first half of 2020. KAS recived nearly 5 million paper-based declarations (including over 2 million PIT-37 declarations).** Processing the data obtained in paper form to digital form is time-consuming and costly for NRA. Often, due to the time-consuming nature of entering all data into NRA systems, this may have a negative impact on the analytical capabilities of tax services.

Idea: Providing a solution using the technology of transforming written and printed text into editable digital form. The solution, on the basis of the expected value in a given field of the declaration, deduces from the paper document and saves the text, data in a digitized form, ensures increased efficiency of automatic conversion from scan to XML file by using data from the National Tax Administration Systems (including dictionaries) in the text recognition process. The final product enables further processing of digital data in other applications and NRA systems.

Basic assumptions

1. Development of software that uses OCR (optical character recognition) technology, i.e. a set of techniques for recognizing characters in a raster graphic file and ICR (Intelligent Character Recognition) technology that allows to read block handwriting for reading tax declarations submitted in paper form.

2. The transformation mechanism should be so sensitive that the employees of NRA will be able to use the currently available (medium-quality) scanning equipment supporting the processing of data from forms submitted in paper versions to create an image of the document.

3. In terms of technical aspects, it will be important to base the proposal on the open source solutions (not currently offered on the market as complete solution).

The aim of the task is primarily to obtain a solution that ensures the link between the scan file (input to the process), the result of the OCR process (intermediate product), the XML file (the result file from the process containing the tax declaration in digital form) and to increase the efficiency of the automatic conversion from scan to XML file through the use of data from the National Revenue Administration systems (including dictionaries) in the text recognition process.

The aim is therefore to provide NRA with a solution that allows the conversion of the paper version of the tax return into an editable digital form (with the correct recognition of text correctness at the highest possible level). The tax declarations subject to conversion are submitted to the NRA in paper form, usually they are filled in by hand, or they are printouts from word processors or programs for filling in tax returns offered by independent entities. The advantage is that each of these documents has the form of a tax return template published by the Ministry of Finance, so such a document contains tax return items according to the scheme required by the Ministry of Finance. The solution can be quickly implemented at NRA due to the established tax return template (layout defined by the tax law).

The solution developed in the form of the DIGITAL_TAX service, after converting the image into text (a scanned document in a pdf, jpg file received electronically) transfers the data to the NRA IT system into the appropriate fields. The algorithms that are part of the system search the recognized content in order to find specific parameters, eg TIN, address, revenue, tax deductible costs, which allows the system to classify expressions in the vicinity of keywords into the appropriate fields. Additionally, the system verifies the data format (e.g. whether a given value is a date), a structure (e.g. TIN), and also analyzes the compliance of the identified data with dictionaries and NRA databases.

The system uses data from the systems of the National Revenue Administration, including data from the taxpayers and payers register, the tax collection system, but also a TERYT address dictionary, a dictionary of public benefit organizations, a dictionary currencies prepared by the National Bank and others.

For the hackathon purpose, the data from the above pointed out systems and dictionaries was simulated with the data included in the XLS file, however, the solution must also be ready to download this data via API (the development of the programming API will be higher scored).

The content recognized by the OCR system that is concentrated on key words has the highest probability of correctness, which, with the simultaneous use of ICR technology, i.e. tools referring to additional data sources, increases the effectiveness of the recognized text. The expected output format of the data is the XML format that complies with the structure of a given declaration - similar to electronic declarations submitted by a taxpayer using the existing solutions (e-Deklaracje system).

Tax returns converted from their paper form to an editable digital form (after OCR and ICR) are sent to the correction process by the user. The solution 'prompts' the user with a text that is not sure that it has been recognized correctly - marking these areas of the text - so that the user can focus on verifying areas marked as uncertain. **At the same time, the solution is to enable the user to modify the entire data (not only the areas indicated as areas for correction).**

Efforts should be made for the user to verify the correctness of the data transformation by correcting the 'proposal' that will be obtained thanks to the solution, if these contain errors reported by the system. The mechanism prompts and signals to the user which items may contain erroneous entries - for verification by the user and possible correction.

The input to the task are prepared by MF / AK:

1.Scanned PIT-37 tax returns.

2. Schemas xsd in versions 25 and 26 of the PIT-37 tax return.

3. Resulting xml files of the given tax return.

4. the data from our systems and dictionaries simulated with the data included in the XLS file,

← The result:

1. GUI for the user (NRA employee), in which he has the opportunity to correct the data, with highlighted areas that are poorly recognized - so that the user can decide on the final values that will be persisted in NRA it systems.

2. XML tax return files, compatible with the supplied XSDs, with the data finally accepted by the user.

In the GUI referred in the point 1 above, the user has access, in the context of a given declaration, to:

- a scan of a given declaration
- generated result file in XML format after OCR compliant with the XSD schema
- the resulting content in the form of a visualized declaration, where the OCR content contains as few areas as possible that could not be recognized correctly, and in the event of difficulties in resolving individual entries, the quality of recognizing the content of the declaration was improved based on data from NRA systems (including taxpayer data, dictionary TERYT, OPP, etc.) and consistent with the structure of the given version of the xsd schema. Quality of data recognition should also be higher by usage of validation criteria described on PIT forms themselves (e.g. that amount in one field should include sum of amounts in two other fields).

Ultimately, the areas that were not recognized properly (despite checking as above) are distinguished from the remaining text - so that the user can correct them and decide on the final values that will be entered into NRA IT systems (i.e. saved in the form of the resulting XML file). Apart from doubtful items, the user must also be able to correct other items of the declaration.

The screen on which the user works to view the results of ICR works is the screen on which the user makes decisions about the content of the data, which will ultimately be recorded in NRA systems - hence the ergonomics of this screen and the output of the decision -making process are of key importance for the evaluation of this solution. The key determinant of the success of this process will be the transformation of paper into a digital version with the lowest possible manual workload for the system operator.

Additionally - it is important that the solution saves (it would save information about the actual operator: the automaton or a specific user) which records were entered by OCR, and which were changed by the user and finally modified.

The final result of the service is an XML file containing data from the scan of the tax declaration (it is important to reproduce the data from the declaration as well as possible). The designed solution is to be transferred to the outside of the application in the resulting XML form (consistent with the structure of the given version of the XSD schema).

Solution development

It is planned that the solution will ultimately cover other tax declarations, hence the service must be both scalable (according to current data in paper form, about 5 million declarations per year are received by NRA) and configurable - so that it can be used with zero or minimal code interference in order to handle other types of declarations (described in XSD

schemas). In terms of technical aspects, it will be important to base the proposal on open source solutions, and in particular on innovative solutions (not currently offered on the market as complete solutions). The architecture of the solution must therefore prepare the solution for:

1. Integration with an external document workflow system / scan source.

2. Integration with an external document workflow system / scan source.

3. Scattering of performance and geographic architecture (scalability).

4. Integration with NRA systems.

5. Security issues (protection of sensitive data, limited access to data, data protection at every stage of processing).

6. Implementation of full accountability of user / system activities in the process.

We ask you to submit your final solutions to our challenge in the form the following files:

1. PROJECT NAME_ Project description.ZIP [MANDATORY]:
   - In English or in Polish
   - No more than 500 words

2. PROJECT NAME_PRESENTATION. ZIP [MANDATORY]:

   - an archive containing all kinds of information materials, descriptions, screenshots, presentations, links to films and other visualizations related to the task, prototype interface etc.
   - presentation (in English or Polish, in PDF only, No more than 10 slides)

3. PROJECT NAME_ MANUAL.ZIP :

Video URL [MANDATORY]:
- Either video of your project or team member explaining the project
- No longer than 60 seconds
- In English or in Polish
-

Demo Link [MANDATORY]:
- Make sure to provide login information

Repository URL [MANDATORY]:
- Everything in one repository (if your project has different modules please put them in separate folders)
- Available to view online
- an archive containing a compiled version of the tool or a link with its location, as well as a description of the steps to run the solution on Windows or Linux in a short time.
- An **automation script** is recommended.

- Framework list, compilers, virtual machines, framework, containers, runtime environment, system requirements, equipment requirements
- Other tools list (for example data base, IT program),
- Operating manual – for system administrator – specific info: configuration how to download files, how to log in, where are OCR files etc.
- Basic user's manual
- Architecture description

4. OTHER Files. ZIP [MANDATORY]:
   - generated result file in XML format after OCR compliant with the XSD schema
   - XML tax return files, compatible with the supplied XSDs, with the data finally accepted by the user. The final result of the service is an XML file containing data from the scan of the tax declaration (it is important to reproduce the data from the declaration as well as possible),