ON-LINE DOCUMENTS CONTENT MANAGEMENT APPLICATIONS

VASILESCU RAMONA VIOLETA
Tibiscus University, Timișoara, Romania
ramonavasilescu@yahoo.com

Abstract:
This paper outlines the steps and technologies used in developing an on-line application server with many desktop clients, and with high power processing for a wide range of input documents to obtain searchable documents on the highest portability standards, PDF and PDF /A.

Key words: *documents management, on-line services, converting documents, PDF*

JEL Classification:

Introduction

Organizing information contained documents was an issue about which we can say that appeared immediately after the invention of writing. The organization was made in the archives of documents: arrive with clay tablets, papyrus archives, document, paper archives, and archives of electronic documents. To find a document in an archive was a need which led to the thinking and implementation of tools to make connection between the document and its potential researcher, as inventory, user archives, archive and catalogue archival advisor (Berciu-Drăghicescu, Iscru, 2005).

Archiving electronic documents raises long-term programs that will be used in the future will have to recognize the internal structure of the document, specific to each type of document. This structure is not visible to the user, the reader has to access the result of a specialized program for reading and playing a certain type of document.

Document management has an increasing importance in the Romanian society. The conclusion in 2009 of the author „Currently, the interest for standard PDF/A and its use in Romania is not developed, even more limiting the announcement that some virtual printers can obtain a PDF file that complies with the PDF/A standard." (Vasilescu, 2009) remained partially valid. In recent years, interest in obtaining the documents in PDF format increased, and, also, the number of articles containing tips on long-term archiving increased.

Material and Methods

The materials used in the research for this paper are research articles, testing of dedicated software and web pages. The study continues other studies about long-time archive and the management of the electronic documents.

Results and discussion

A first step in the research for this paper was to determine interest for long-term archiving in Romania. This interest is growing and is demonstrated by the appearance of several articles devoted to long-term archiving. These articles bring to public attention in the form arguments for the electronic document management (safety, search and access more efficient, economic and ecological benefits, social responsibility) (DMS ONE, 2011). Meanwhile, in Romania, the associated services with long-time archived are developing and diversifying. Services are considering both paper documents and electronic documents: document storage, records management, secure destruction, scanning and digitization, electronic archiving.

Archiving electronic documents on long term is increasingly discussed topic in the online environment in Romania. The discussion covers both the law governing the issuance and electronic storage of invoices, receipts and tax bills and issues that arise from the informational long-term archiving.

Programs dedicated accounting activities have tools to save documents that were printed and electronic format. The files can be obtained in TIFF or PDF. Over many years of archival documents through the use of programs and / or different instruments in any company there will be an archive file that contains at least two formats: PDF and TIFF. Also, the archive may contain emails with or without attachments saved in specific file types (ex. MSG or EML). Inevitably, should be considered the question of how we can "flatten" the archive to contain documents simultaneously and content of documents to be accessed for search (which requires the application of OCR technology).

Applications for use of OCR technology can be grouped as follows:

A1. Applications of technology providers;

A2. Applications created by software developers based on components supplied by OCR technology creators.

Each of these two main groups have advantages and disadvantages, such as:

- Applications of Group A1 shows confidence because they are results of OCR technology providers work (ABBY FineReader, for example);
- Applications of group A2 may be more affordable due to the fact that developers do not need to invest time and effort in developing OCR technology itself.

Applications with complex interface working directly on the text of a document open in a window are not useful if desired massive conversion of documents containing searchable for performing the same steps that involve time-consuming for each document, even if they are running in semi-automatic manner (i.e. the user selects an option to work):

- Open the document;
- Specify the area to be used for OCR;
- Saving the document.

If massive processing, it is recommended programs that automates almost all the steps described above.

In this paper we analyse a solution through a program dedicated to obtaining PDF or PDF/A format from image files, like TIFF and JPG. The program is available on the May Computer company's website at http://www.ocrserver.at. Drop OCR program uses a Web service to send files to be converted to a dedicated server in applying OCR technology.

The AutoOCR server of May Computer Company is a powerful server and provides an interface of type SOAP / REST. To access this server, an Internet connection is required.

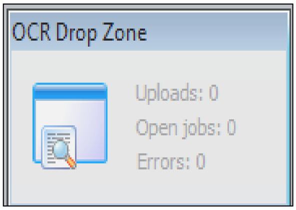
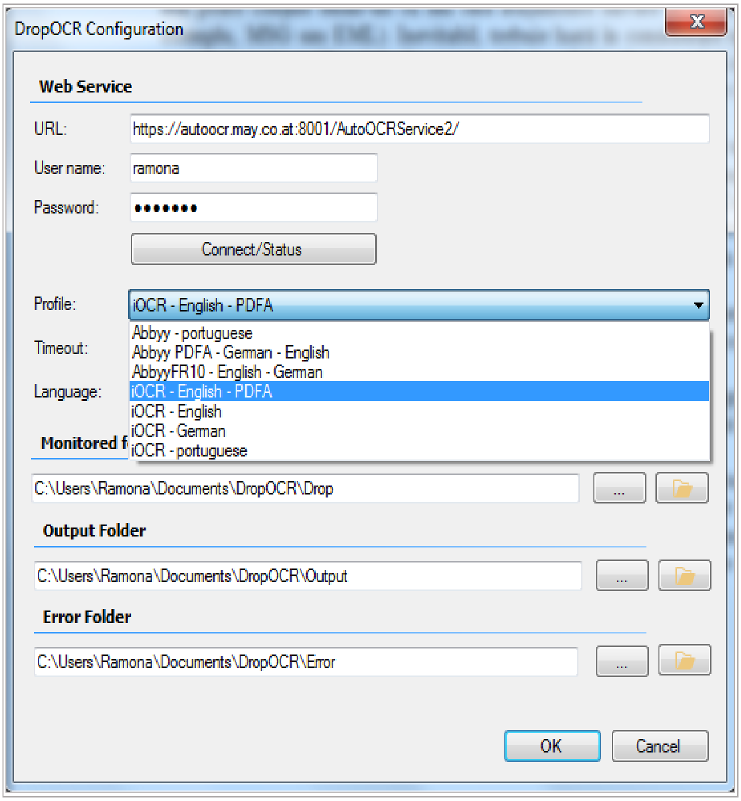


Figure 1. The **OCR Drop Zone** window

Working with the program is simple. After installing and launching it in execution, it is noted the appearance of a small window in the lower right side (see Figure 1) **Drop Zone OCR** name.

After installation, the user must configure some specific options (see Figure 2 for an example), such as the server address, username, password.

As shown in the Figure 2, the user can select a profile. In this application, a profile is a name given to a set of server work options. From the list of profiles, the user can choose and can deduce which profile is better to be applied according the language of the archived document or documents.



DropOCR Configuration

Web Service

URL: https://autoocr.may.co.at:8001/AutoOCRService2/

User name: ramona

Password: •••••••

Profile: iOCR - English - PDFA

Monitored f

C:\Users\Ramona\Documents\DropOCR\Drop

Output Folder

C:\Users\Ramona\Documents\DropOCR\Output

Error Folder

C:\Users\Ramona\Documents\DropOCR\Error

Figure 2. Configuration options for **DropOCR**

In the configuration window, the user specifies an input folder which will be constantly monitored as follows: in case of a file it will be automatically sent to the server, the result obtained from the server being submitted:

- If valid, in the output folder,

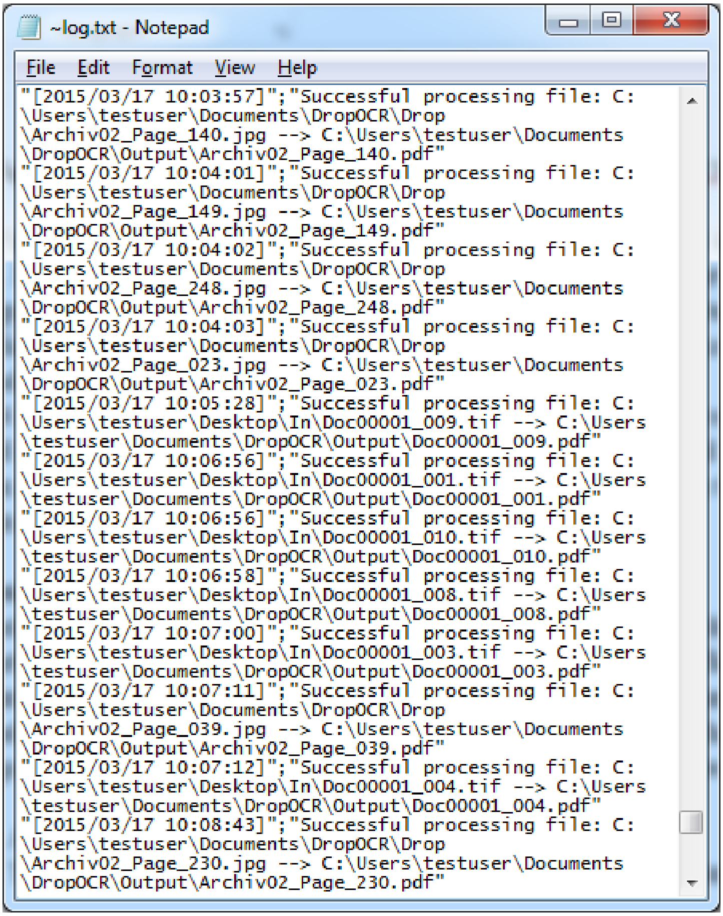
- If it is wrong in the errors folder.

The window shown in Figure 1, works similarly except that the monitored folder files that are processed are those that were "pulled" into this window by a drag & drop operation.

Figure 3 shows:

- The final part of the information about the result of the conversion of an archive consisting of more than 200 files, each file representing a document page; these files were taken automatically from monitored folder, for instance: Archiv02_Page_023, Archiv02_Page_248, Archiv02_Page_039;

- The information about the result of the conversion of files sent by drag & drop; for instance: Doc00001_009, Doc00001_001, Doc00001_010.



```
"[2015/03/17 10:03:57]";"Successful processing file: C:
\Users\testuser\Documents\DropOCR\Drop
\Archiv02_Page_140.jpg --> C:\Users\testuser\Documents
\DropOCR\Output\Archiv02_Page_140.pdf"
"[2015/03/17 10:04:01]";"Successful processing file: C:
\Users\testuser\Documents\DropOCR\Drop
\Archiv02_Page_149.jpg --> C:\Users\testuser\Documents
\DropOCR\Output\Archiv02_Page_149.pdf"
"[2015/03/17 10:04:02]";"Successful processing file: C:
\Users\testuser\Documents\DropOCR\Drop
\Archiv02_Page_248.jpg --> C:\Users\testuser\Documents
\DropOCR\Output\Archiv02_Page_248.pdf"
"[2015/03/17 10:04:03]";"Successful processing file: C:
\Users\testuser\Documents\DropOCR\Drop
\Archiv02_Page_023.jpg --> C:\Users\testuser\Documents
\DropOCR\Output\Archiv02_Page_023.pdf"
"[2015/03/17 10:05:28]";"Successful processing file: C:
\Users\testuser\Desktop\In\Doc00001_009.tif --> C:\Users
\testuser\Documents\DropOCR\Output\Doc00001_009.pdf"
"[2015/03/17 10:06:56]";"Successful processing file: C:
\Users\testuser\Desktop\In\Doc00001_001.tif --> C:\Users
\testuser\Documents\DropOCR\Output\Doc00001_001.pdf"
"[2015/03/17 10:06:56]";"Successful processing file: C:
\Users\testuser\Desktop\In\Doc00001_010.tif --> C:\Users
\testuser\Documents\DropOCR\Output\Doc00001_010.pdf"
"[2015/03/17 10:06:58]";"Successful processing file: C:
\Users\testuser\Desktop\In\Doc00001_008.tif --> C:\Users
\testuser\Documents\DropOCR\Output\Doc00001_008.pdf"
"[2015/03/17 10:07:00]";"Successful processing file: C:
\Users\testuser\Desktop\In\Doc00001_003.tif --> C:\Users
\testuser\Documents\DropOCR\Output\Doc00001_003.pdf"
"[2015/03/17 10:07:11]";"Successful processing file: C:
\Users\testuser\Documents\DropOCR\Drop
\Archiv02_Page_039.jpg --> C:\Users\testuser\Documents
\DropOCR\Output\Archiv02_Page_039.pdf"
"[2015/03/17 10:07:12]";"Successful processing file: C:
\Users\testuser\Desktop\In\Doc00001_004.tif --> C:\Users
\testuser\Documents\DropOCR\Output\Doc00001_004.pdf"
"[2015/03/17 10:08:43]";"Successful processing file: C:
\Users\testuser\Documents\DropOCR\Drop
\Archiv02_Page_230.jpg --> C:\Users\testuser\Documents
\DropOCR\Output\Archiv02_Page_230.pdf"
```

Figure 3. Information about the results of conversions via the DropOCR

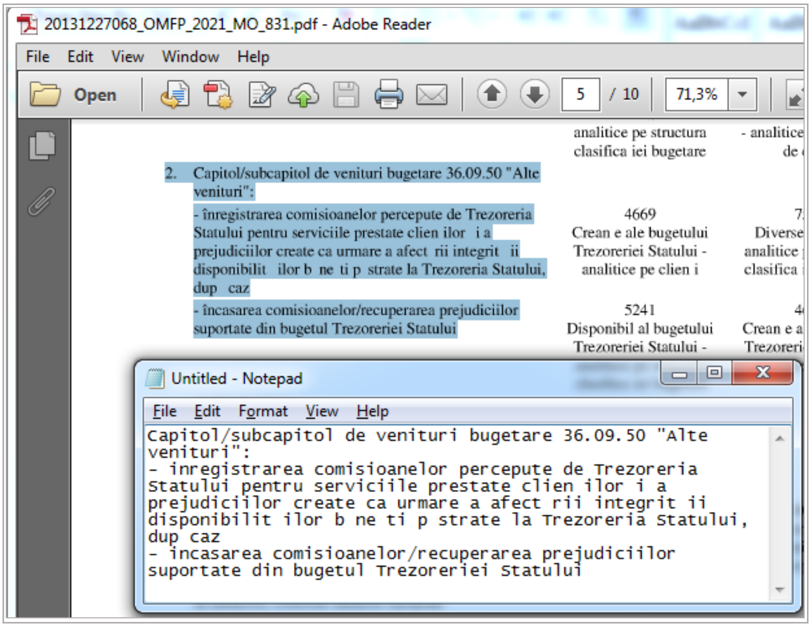
In the Figure 3 we can observe that the result files are obtained in an order that seems random. In fact, each file is obtained depending on when the server finishes its processing. Apparently, this may be a disadvantage of this application but bear in mind that, if the application is used for massive conversions of old archives, counts the result to be correct and complete.

Figure 4 shows as an example a PDF file obtained after the conversion of a TIF file. It is noted that the text can be selected, copied to clipboard and thence copied elsewhere, even for scanned images. Therefore, we can conclude that the text obtained is useful for use in content searches or other processing.

Conclusions

The need for archiving and "uniformity" archives of electronic documents is an increasing target not only in Romania. Automated conversion of the different formatted documents in an archive containing documents of the same type and searchable content bring a strong advantage in the life of a company that has clear benefits that contribute to personal savings, time, and material resources. Time resource to which we refer has two components:

- Actual time used to convert documents,
- Time in the future used to search by the documents' content.


20131227068_OMFP_2021_MO_831.pdf - Adobe Reader
File Edit View Window Help
Open
5 / 10
71,3%
Untitled - Notepad
File Edit Format View Help

analitice pe structura clasifica iei bugetare | - analitice de

2. Capitol/subcapitol de venituri bugetare 36.09.50 "Alte venituri":

- înregistrarea comisioanelor percepute de Trezoreria Statului pentru serviciile prestate clien ilor i a prejudiciilor create ca urmare a afect rii integrit ii disponibilit ilor b ne ti p strate la Trezoreria Statului, dup caz | 4669 Crean e ale bugetului Trezoreriei Statului - analitice pe clien i | 7 Diverse analitice clasifica

- încasarea comisioanelor/recuperarea prejudiciilor suportate din bugetul Trezoreriei Statului | 5241 Disponibil al bugetului Trezoreriei Statului - | 4 Crean e a Trezoreri

```
Capitol/subcapitol de venituri bugetare 36.09.50 "Alte
venituri":
- inregistrarea comisioanelor percepute de Trezoreria
Statului pentru serviciile prestate clien ilor i a
prejudiciilor create ca urmare a afect rii integrit ii
disponibilit ilor b ne ti p strate la Trezoreria Statului,
dup caz
- incasarea comisioanelor/recuperarea prejudiciilor
suportate din bugetul Trezoreriei Statului
```

Figure 4. Result obtained by OCR

We believe that choosing a dedicated application for reorganization old archives and papers on new principles (such as the possibility to construct searchable content archived files) we must consider the level of automation of processing to be applied files. Large volumes of archived files in different formats, requires a high degree of automation. Drop Zone is an example of application that proofs the truth of this statement.

References

Berciu-Draghicescu, Adina, Iscru G.D. (2005) Introducere în știința istorică și în științele auxiliare ale istoriei. Surse info-documentare, Editura Universității din București

Vasilescu, Ramona (2009) PDF/A standard for long term archiving, Annals. Computer Science Series. 7th Tome 1st Fasc

*** DMS ONE (2011) Patru argumente pentru managementul documentelor în formă electronică, 25.01.2011, article available electronically at http://www.dmsone.ro/Professional-articles/Patru-argumente-pentru-managementul-documentelor--n-form-electronic-