SWEN-440

https://www.se.rit.edu/~swen-440

Project 2 – Processing Services Overview

SW System Requirements Engineering and Architecture

Contents

[Summary 2](#_Toc73016697)

[Document Repository (DMService) 3](#_Toc73016698)

[OCR service (OCRService) 4](#_Toc73016699)

[Parser (Text Extraction) Service 5](#_Toc73016700)

[OCR FTP Service 6](#_Toc73016701)

[General Sequence 6](#_Toc73016702)

[DM FTP Service 7](#_Toc73016703)

# Summary

This document provides information on the main web services available for building the Process System. There is also an FTP service for batch processing (described later in the document)

The web services are:

* Document Repository (DMService)
* OCR Service (OCRService)
* Parser (Text extraction) service

Each is described below, and there is information at the WebServices home page.

<http://seappserver1.rit.edu/FormsServices>

# Document Repository (DMService)

The document repository is a central location for all the received documents.

These are placed in folders that are accessible in two ways

1. Via Ftp
2. Via Http Download

If using Ftp, a standard Ftp Client can be used. Contact the system administrator for a login

If using Http download, two APIs are provided (files are encoded using MIME in the body of the HTTP transaction)

* List files
* Get file

See the URL @ <http://seappserver1.rit.edu/FormsServices> for details

# OCR service (OCRService)

The OCR Service takes an image file as input, and returns the text content of the file.

It can only take a single page image to process. Currently only .png or .jpg files are supported.

The OCR service has two modes of operation

* Async processing
  + In this mode, files are uploaded to a designated folder (folder and subfolder creation are handled by the service). As files are uploaded the OCR service will start processing the files, and place a text file (with the same name as the original file, with a .txt extension) in the subfolder. The client is responsible for retrieving the output (.txt) files. In this mode, you may submit multiple files for processing and retrieve the results later. This is to allow batch processing operations.
* Synchronous/ Single file processing
  + A REST API is provided to upload a file, and receive the name of the output file in the JSON return data. The client is responsible for retrieving the output (txt) file. See the website at <http://seappserver1.rit.edu/FormsServices> for API information

There are two type of OCR processing engines available. They are referred to as ‘standard’ (std) and ‘professional’ (pro).

* The standard engine is bundled with the OCR application and no incremental pricing is required.
* The professional OCR engine has incremental cost. There is an additional charge of $1.00 per document processed, or a flat monthly fee of $1,000 per month per server instance.

# Parser (Text Extraction) Service

This service will take a text file upload, and return the key word extractions in a JSON result. See the website at <http://seappserver1.rit.edu/FormsServices> for details.

# OCR FTP Service

An alternate to the Web API mechanism, is to use the Batch Processing capabilities of the system. This allows users to deposit files using FTP, and having the files be processed by the OCR Service in bulk. Files can be retrieved when ready.

NOTE: To use the service, see your instructor to have an FTP account/ folder assigned to your team.

## General Sequence

Send FTP File to <ftp\_server>/xxxx

* xxx is your assigned FTP folder based on your assigned accountId. When you login using the assigned accountId, you will be placed in your appropriate home directory (xxxx)

Monitor output folder <ftp\_server>/xxxx/completed

* Once the file(s) are in the xxxx directory, OCR processing will start. The processed files will be placed in the ‘completed’ directory. Use FTP to monitor that directory to retrieve the results of processing. The processed file (in the ‘completed’ folder) will be the same name as the original file, but with a .txt extension. E.g. Input file was “MyFile.png”, Completed file will be “MyFile.txt” in the ‘Completed’ directory.

Retrieve .txt files from the ftp\_server

* Once the file is located in the ftp completed directory, just retrieve the file, also using FTP.

# DM FTP Service

In addition to RESTful API listing and retrieval of documents, the FTP protocol can also be used for accessing documents.

* Document Management services provides a mirror site
  + The FTP server retains a ‘mirror’ of the DM site. i.e. the application .png files are also available from the FTP site.
  + You may retrieve the list of the files using a ‘dm’ compatible account, and also retrieve the files for further processing using standard FTP commands.
* Account information:
  + The account credentials are:
    - Username: ftp\_dm
    - Password: ftp\_dm\_password
* Using this login, if you list files, you will get a full list of all available files
* You may then use ftp to retrieve specified files.