ELO OCR Textreader für Abbyy FineReader

[Stand: 05.05.2010]

Inhalt

1 Architekturübersicht

Die folgenden Ausführungen gehen davon aus, dass Sie mit dem ELOenterprise Volltextdienst und den Abläufen im Volltextdienst vertraut sind. Wenn das nicht der Fall ist, sollten Sie erst diese Anleitungen lesen. Andernfalls werden viele Erläuterungen aus diesem Dokument unverständlich bleiben.

Der ELO OCR Textreader besteht aus zwei Teilen:

a) dem Service-Modul für die Windows Service Schnittstelle
b) dem Übersetzungsmodul

Das Service-Modul bedient die Windows Service Schnittstelle. Hierzu gehören der Start der Übersetzungsmodule beim Dienststart und das Beenden der Übersetzungsmodule bei Dienstende. Da die Kommunikation zwischen den Modulen asynchron über die Registry erfolgt, ergibt sich beim Beenden eines langlaufenden OCR Prozesses keine Blockierung des Windows Service Managers.

Das Übersetzungsmodul überwacht ein Eingabeverzeichnis auf TIFF-Dateien und führt dann eine OCR Konvertierung in das Ausgabeverzeichnis durch. Dieses Modul ist vollkommen unabhängig vom Service-Modul und kann mehrfach gestartet werden (z.B. auf Multiprozessor Servern).

Beim Dokumentenfluss gibt es zwei unterschiedliche Szenarien, je nach Anzahl der verwendeten OCR Threads. Diese werden im Folgenden erläutert.

1.1 Dokumentenfluss mit einem OCR Thread

Wenn Sie nur einen OCR Thread verwenden, dann können Sie das Ausgabeverzeichnis für Tiff-Dateien der ELOenterprise Volltextengine (FT_OUT) direkt als Eingabeverzeichnis für das Übersetzungsmodul (ORC_IN) verwenden. Das Ausgangsverzeichnis des Übersetzungsmoduls (OCR_OUT) ist gleich dem Eingangsverzeichnis der Volltextengine (FT_IN).

Das Dokument läuft dann folgenden Weg:

1. Der Volltextdienst findet ein neues Dokument und speichert es im FT_OUT.
2. Da FT_OUT gleich OCR_IN ist, findet das Übersetzungsmodul diese Datei in seinem Eingangsverzeichnis vor und startet den OCR Vorgang.
3. Wenn der OCR Vorgang abgeschlossen ist, legt das Übersetzungsmodul eine Textdatei im OCR_OUT Verzeichnis ab.
4. Da OCR_OUT gleich FT_IN ist, wird diese Textdatei vom Volltextdienst aufgegriffen und in die Datenbank übernommen.

Wenn Sie den OCR Textreader auf einem Server gemeinsam mit anderen Programmen betreiben, dann sollten Sie nur einen OCR Thread aufsetzen. In diesem Fall ist es ja nicht gewünscht, dass der OCR Vorgang die komplette Serverleistung in Anspruch nimmt. Zusätzlich sollten Sie die Prozesspriorität auf „NIEDRIG“ stellen (Achtung: in diesem Fall sollten Sie keine aufwendigen Bildschirmschoner einsetzten, da sie die gleiche Priorität wie der OCR Vorgang haben).

1.2 Dokumentenfluss mit mehreren OCR Threads

Etwas komplizierter wird die Situation, wenn Sie mehrere OCR Threads benötigen. Das ist z.B. dann der Fall, wenn Sie einen dezidierten OCR Server (evtl. sogar Multiprozessor) einsetzen und die Maschine optimal auslasten wollen. In diesem Fall sollten Sie pro CPU einen eigenen OCR Thread einplanen, bei SMT CPUs sogar 2 OCR Threads pro CPU. In dieser Konfiguration erreichen Sie eine Gesamtauslastung von nahe 100% und somit eine maximale OCR Leistung.

Da die OCR Prozesse untereinander nicht kommunizieren, können diese nicht auf ein gemeinsames Eingangsverzeichnis zugreifen. Das würde dazu führen, dass alle Threads die gleiche (erste) Datei bearbeiten und sich gegenseitig stören. Sie müssen stattdessen eine Konfiguration erstellen, in der jeder OCR Thread ein eigenes Eingangsverzeichnis hat.

Das Ausgangsverzeichnis dagegen kann ein gemeinsames Verzeichnis sein, im Normalfall das Eingangsverzeichnis für den Volltextdienst. Zur Verteilung der Dokumente aus dem Ausgangsverzeichnis des Volltextdienstes (FT_OUT) auf die Eingangsverzeichnisse der OCR Threads (OCR_IN1, OCR_IN2 ...) gibt es im Service-Modul eine Verteilroutine. Dies prüft in regelmäßigen Zyklen die OCR_IN* Verzeichnisse ab. Wenn dort weniger als 10 Dateien liegen, kopiert es 20 weitere Dateien aus dem SplitInput (sollte auf FT_OUT eingestellt werden) Verzeichnis. Über diese dynamische Verteilung soll sichergestellt werden, dann nicht einige OCR Threads mangels neuer Dokumente still stehen und andere Threads noch eine Vielzahl nicht abgearbeiteter Dokumente haben.

Ein Dokument läuft dann folgenden Weg:

1. Der Volltextdienst findet ein neues Dokument und speichert es in FT_OUT
2. Sobald das Service-Modul im Split Thread feststellt, dass ein Übersetzungsprozess neue Dokumente benötigt, wird das Dokument in das entsprechende OCR_IN* Verzeichnis verschoben.
3. Das Übersetzungsmodul greift das Dokument auf und führt eine OCR Analyse durch.
4. Die erzeugte Textdatei wird im OCR_OUT Verzeichnis abgespeichert.
5. Da OCR_OUT gleich FT_IN ist, findet der Volltextdienst die neue Textdatei und übernimmt diese in seine Datenbank.

Das SplitInput Verzeichnis sollte standardgemäß für TIFF Dateien verwendet werden. Für JPG Dateien gibt es das SplitInput2 Verzeichnis welches sich analog dazu verhält.

Achtung: der Transfer vom SplitInput (normalerweise FT_OUT) Verzeichnis in die OCR_IN* Verzeichnisse muss eine „atomare" Operation sein. Das können Sie nur sicherstellen, indem die FT_OUT und OCR_IN* Verzeichnisse auf der gleichen Partition im gleichen logischen Laufwerk liegen. Andernfalls wird der MOVE Befehl von Windows automatisch durch ein COPY + DELETE simuliert, was dazu führen kann, dass die OCR Software mitunter unvollständige Tiff-Dateien erhält. Diese würden dann zu Fehlermeldungen und unvollständig erkannten Dokumenten in der Volltextsuche führen.

2 Installation

Im Abschnitt „Beispiele für die Konfiguration“ wird eine Verzeichnisstruktur für den OCR Textreader vorgeschlagen. Für spezielle Belange können Sie davon abweichen, im Normalfall sollten Sie diese Struktur aber im Kern für Ihre eigene Installation übernehmen. Dabei wird auch ein „Pgm“ Verzeichnis für die Programme angelegt.

1. Führen Sie von der CD das Fulltext-OCR Setup aus.
2. Falls Sie das Setup in Form einer Zip-Datei erhalten haben, packen Sie es in einem temporären Verzeichnis aus.
3. Starten Sie das Setup-Programm
4. Geben Sie dort das TIF-Ausgangsverzeichnis des Volltextdienstes an, hier holt sich der OCR Textreader seine Dateien für die Analyse ab.
5. Geben Sie das TXT-Rückgabeverzeichnis des Volltextdienstes an, hier legt der OCR Textreader die erkannten Texte ab.
6. Geben Sie das Arbeitsverzeichnis für den OCR Textreader an. Hier legt das Setupprogramm eine Unterstruktur mit den benötigten Verzeichnissen für die temporären Dateien des Analysevorgangs an.
7. Geben Sie das BIN Verzeichnis des Abbyy FineReader OCR Verzeichnisses an. In diesem Verzeichnis muss die Datei FREngine.DLL vorhanden sein. Da diese von dem Textreader verwendet wird, führt eine Fehleingabe dazu, dass der OCR Prozess nicht durchgeführt werden kann.
8. Geben Sie die Anzahl der Textreader-Threads an. Falls der OCR Vorgang gemeinsam mit anderen Komponenten auf einem Server läuft, sollten Sie hier nur einen Thread angeben, da es nicht erwünscht ist, die komplette Rechenleistung mit der OCR Analyse zu verbrauchen. Wenn Sie einen eigenen Server nur für den Textreader haben, dann sollten Sie hier pro Prozessor einen oder zwei Threads einplanen.
9. Das ReloadInterval sollten Sie im Normalfall auf den Standardwert von 24 Stunden belassen.
10. Stellen Sie die Thread Priorität auf „low“, wenn der Textreader keine eigene Maschine hat.
11. Geben Sie das Windows Konto für den Textreader Prozess an. Es wichtig, dass dieses Konto vollen Zugriff auf die eingestellten Verzeichnisse besitzt (Achtung: Dienste haben im Allgemeinen keinen Zugriff auf manuell gemappte Verzeichnisse!). Dieses Konto benötigt keine besonderen Rechte, insbesondere keine Administrationsrechte. Falls Sie „aus Bequemlichkeit“ das lokale Systemkonto verwenden, sollten Sie bedenken, dass dieses nur Zugriff auf lokale Resourcen besitzt (hier aber sehr weitreichende Berechtigungen).
12. **Wenn Sie den Service nicht unter dem Systemkonto betreiben, müssen Sie darauf achten, dass das gewählte Dienstkonto Vollzugriff auf den "HKLM\Software\ELO Digital" Registry Pfad besitzt. Zudem muss er natürlich auch im Dateisystem Vollzugriff auf alle beteiligen Verzeichnisse (Input, Output, Temp…) besitzen. Im Fehlerfall kann das zu nur schwer erkennbaren Fehlfunktionen führen.**

3 Konfiguration

3.1 Service-Modul

Die Konfiguration des Service-Moduls wird in der Registry unter HKLM\Software\ELO Digital\Fulltext hinterlegt. Beachten Sie bitte, dass dieses Modul in diesem Registryzweig Schreibberechtigung benötigt, da die Kommunikation zu den Übersetzungsmodulen über die Registry erfolgt.

Name	Funktion	Beispiel
LowPriority	DWORD, 0 oder 1. Legt fest, ob die Übersetzungsmodule mit normaler (0) oder niedriger (1) Priorität gestartet werden. Default 0: normale Priorität	1
OCRThreads	DWORD, 0..9. Legt die Anzahl der zu startenden Übersetzungsmodule fest. Default 1: es wird ein Übersetzungsmodul gestartet	2
ReloadInterval	DWORD. Legt die Zahl von Sekunden fest bis das Service-Modul die laufenden Übersetzungsmodule beendet und neu startet. Ziel dieser Option ist es, mögliche Memory Leaks in der Übersetzungssoftware durch ein regelmäßigen Neustart der Module auszugleichen. Da bei einem Reload alle Übersetzungsmodule beendet werden, ist sichergestellt, dass das Betriebssystem alle verlorenen Resourcen wieder freigeben kann. Default: 86400, d.h. es wird ein Neustart pro Tag ausgeführt.	86400
ServiceReport	Zeichenkette. Dieser Eintrag enthält	C:\temp\ServiceReport.txt

	den Pfad und Dateinamen für die Reportausgabe des Service-Moduls. Falls dieser Eintrag nicht existiert oder leer ist, wird kein Report geschrieben. Ein fehlerhafter Eintrag kann zu nicht reproduzierbaren Fehlern oder Schutzverletzungen führen. Default: kein Eintrag	
SplitInput	Zeichenkette. Der ELOenterprise Volltextdienst exportiert die TIFF Dateien in ein Ausgabeverzeichnis. Falls Sie mehrere Übersetzungsmodule verwenden wollen, benötigt jedes Modul ein eigenes Eingabeverzeichnis. Über die Split Funktion können Sie das Service-Modul dazu anweisen, den Inhalt des Volltext-Ausgabeverzeichnises auf die verschiedenen Eingabeverzeichnisse zu verteilen. Dazu geben Sie hier den Pfad des Volltext-Ausgabeverzeichnisses für Tiff Dokumente mit einem abschließenden Backslash an. Falls Sie nur mit einem Übersetzungsmodul arbeiten, sollten Sie diesen Eintrag leer lassen und das Volltext-Ausgabeverzeichnis direkt in der Konfiguration des Übersetzungsmoduls angeben. Default: kein Eintrag	C:\FT\TIF\
SplitInput2	Zeichenkette. Wie SplitInput, aber für ein zweites Verzeichnis für JPG Dateien. Default: kein Eintrag	C:\FT\JPG\

Die Zahl der zu verwendenden Übersetzungsmodule richtet sich nach der Aufgabe und der zulässigen Serverlast.

Falls Sie den OCR Textreader auf dem normalen ELOenterprise Server mit laufen lassen, dann ist es nicht erwünscht, dass der Übersetzungsprozess zu viel Rechenleistung in Anspruch nimmt. In diesem Fall sollten Sie nur mit einem Übersetzungsprozess, der auf niedrige Priorität eingestellt wird, arbeiten.

Wenn Sie einen eigenen Computer für den Textreader zur Verfügung haben, dann sollte dieser möglichst gut durch die Übersetzungsprozesse ausgelastet werden. Insbesondere wenn dieser über mehrere CPUs verfügt, wäre der Einsatz eines einzigen Übersetzungsmoduls eine Verschwendung von Ressourcen. Hier sollten Sie so viele Übersetzungsmodule konfigurieren wie CPUs vorhanden sind, falls es sich um CPUs mit SMP Fähigkeit handelt, ist sogar die doppelte Anzahl von Prozessen sinnvoll.

Das ReloadInterval bestimmt die Verweildauer des OCR Moduls im Speicher. Jeweils nach Ablauf der eingestellten Zeit (in Sekunden, minimal 600) werden alle aktuell aktiven Übersetzungsmodule gestoppt und anschließend neu gestartet. Durch den vollständigen Stop aller Programme, die die OCR Software nutzen, hat das Betriebssystem die Möglichkeit, alle verwendeten Ressourcen wieder frei zu geben. Da beim Stop die Übersetzungsmodule nicht einfach abgebrochen werden, sondern nur aufgefordert werden, nach Abschluss der laufenden Übersetzung sich selber zu beenden, kann dieser Vorgang mitunter auch eine längere Zeit in Anspruch nehmen. Das ist kein Fehler, dieses Verhalten ist in der Neustartfunktion eingeplant.

Falls ein Übersetzungsprozess durch eine Schutzverletzung oder einen externen Eingriff abgestürzt ist, wird dieser Zustand erkannt und automatisch ein neuer Übersetzungsprozess gestartet.

3.2 Übersetzungsmodul

Die Übersetzungsmodule werden von 1 an aufsteigend durchnummeriert. Wenn Sie also 4 OCR Threads im Service-Modul definiert haben, dann gibt es die Übersetzungsmodule 1,2,3 und 4.

Die Konfiguration des Moduls <n> findet sich in der Registry unter HKLM\Software\ELO Digital\Fulltext.<n> (also z.B. für das Modul 1 unter ... \Fulltext.1). Auch das Übersetzungsmodul benötigt auf diesen Registryzweig vollen Schreib/Lesezugriff, andernfalls läuft die Quittierung des Systemstops nicht korrekt.

Diese Werte sollten auch im Service Modul hinterlegt werden, da sie dann beim Anlegen eines neuen Übersetzungsprozesses als voreingestellte Werte verwendet werden.

Name	Funktion	Beispiel
Broken	Zeichenkette: Verzeichnisname für defek-	D:\FULLTEXT\BROKEN\

	te Tiff-Dateien. Wenn der OCR Prozess durch eine Fehlermeldung oder eine Schutzverletzung abgebrochen wird, dann wird das aktuelle Dokument in dieses Verzeichnis verschoben. Sie haben hier die Möglichkeit zur Analyse, welche Dokumente nicht abgearbeitet werden konnten. Falls Sie diesen Eintrag leer lassen, werden diese Dateien einfach gelöscht, Sie erhalten dann keine Rückmeldung über fehlgeschlagene OCR Vorgänge. Default: keiner	
FineReader	Zeichenkette: Zugriffspfad auf die OCR Engine, mit Dateiname der DLL Schnittstelle (FREngine.DLL im Bin-Verzeichnis des ABBYY FineReaders). Default: keiner, Eingabe notwendig	C:\Programme\ABBYY FineReader Engine 7.0\Bin\FREngine.dll
InputDir InputDir2 InputDir3	Zeichenkette: Eingangsverzeichnis für die Tiff-Dateien. Dieses Verzeichnis darf nur von einem Übersetzungsmodul verwendet werden. Wenn Sie mit mehreren Übersetzungsmodulen arbeiten, dann müssen Sie auch mehrere Eingangsverzeichnisse anlegen. Dieser Eintrag muss auf ein bestehendes Verzeichnis zeigen und einen abschließenden Backslash enthalten. Default: keiner, Eingabe notwendig Optional kann der Volltextprozess auch weitere Verzeichnisse überwachen. Hierfür werden dann die Registryeinträge InputDir2 und InputDir3 verwendet. Diese Einstellung ist dann sinnvoll, wenn PDF Scandokumente mit eingebetteten JPG Scanseiten ausgewertet werden sollen.	D:\FULLTEXT\RDR1\TIF\ D:\FULLTEXT\RDR1\JPG\

Languages	Zeichenkette: Voreingestellte Sprachen für den FineReader. Diese bestimmen die verwendeten Wörterbücher sowie den erkannten Zeichensatz. Default: german	german,french,english
LogFile	Zeichenkette: Log-Datei für Ausgabemeldungen des Abbyy FineReaders mit Pfad und Dateiname. Dieser Eintrag ist unbedingt notwendig, wenn er fehlt, versucht der OCR Prozess diese Meldungen per MessageBox auszugeben. Das ist bei einem Serverprozess extrem nachteilig, bei einem Dienst führt es zu erheblichen Fehlfunktionen welche bis zum Hängen des Prozesses führen können. Default: keiner, Eingabe notwendig	D:\FULLTEXT\LogFile1.TXT D:\FULLTEXT\LogFile2.TXT
OutputDir	Zeichenkette: Ausgabeverzeichnis für die erstellten Textdateien. Dieses Verzeichnis ist normalerweise das TXT-Rückgabeverzeichnis vom ELOenterprise Volltextmodul. Dieser Eintrag muss auf ein bestehendes Verzeichnis zeigen und einen abschließenden Backslash enthalten. Default: keiner, Eingabe notwendig	D:\FULLTEXT\TXT\
Pages	DWORD: Dieser Eintrag wird vom Übersetzungsmodul erstellt und gefüllt. Er enthält die Zahl der verarbeiteten Seiten (nicht Dokumente). Default: keiner, optional, wird automatisch erstellt	0
Report	Zeichenkette: Pfad und Dateiname für die Reportdatei des Übersetzungsmoduls. Falls mehrere Module aktiv sind, benötig jedes eine eigene Reportdatei. Wenn keine Reportdatei angegeben wird, wird	D:\FULLTEXT\Report1.TXT D:\FULLTEXT\Report2.TXT

	auch kein Report erstellt. Default: kein Report	
Stop	DWORD, 0 oder 1: mit einer 1 fordert das Service-Modul eine Beendigung dieses Übersetzungsmoduls an. Dabei gibt es keinen harten Abbruch sondern das Übersetzungsmodul beendet sich erst nach der vollständigen Abarbeitung des aktuellen Dokuments. Der Abbruch erfolgt dabei über ein Handshake: das Service-Modul setzt hier eine 1 um ein Abbruch des Prozesses anzufordern. Sobald das Übersetzungsmodul die Verarbeitung des aktuellen Dokuments beendet hat, setzt es diesen Wert zur Quittierung auf 0 zurück und beendet sich. Default: 0	0
TempDir	Zeichenkette: Der OCR Prozess muss zur Analyse der Seiten eine Aufteilung des Dokuments in Einzeldateien vornehmen. Weiterhin wird während der Analyse zu jeder Seite eine temporäre Layout Datei erzeugt. Diese temporären Dateien werden am Ende des OCR Vorgangs automatisch gelöscht. Obwohl sich aus technischer Sicht mehrere Prozesse ein TempDir teilen könnten, sollte man aus organisatorischer Sicht für jedes Übersetzungsmodul ein eigenes Verzeichnis einrichten. Falls es zu Problemen kommt, kann man die fehlerhaften Dateien schneller einem Prozess zuordnen. Dieser Eintrag muss auf ein bestehendes Verzeichnis zeigen und einen abschließenden Backslash enthalten. Default: keiner, Eingabe notwendig	D:\FULLTEXT\RDR1\Temp\ D:\FULLTEXT\RDR2\Temp\

	ACHTUNG: das Verzeichnis wird bei jedem Programmstart automatisch geleert um hängende Dateien aus vorangehenden Programmläufen zu entfernen. Wenn Sie hier eigene Daten ablegen, können diese verloren gehen.	
WatchDog	DWORD: Wartezeit des Watchdog Timers bis zum Abbruch eines festhängenden OCR Prozesses. Dieser Wert richtet sich nach der maximalen Bearbeitungszeit einer komplexen Seite (nicht des kompletten Dokuments). Bei handelsüblicher Hardware mit normaler Belastung ist hier ein Wert zwischen 30 und 200 Sekunden sinnvoll. Wenn dieser Wert zu kurz angegeben wird, dann kann es vorkommen, dass Seiten verworfen werden, die mit etwas mehr Zeit doch analysierbar wären. Ist diese Zeit zu lang, wird ein festhängendes Dokument den OCR Textreader unnötig lang aufhalten und es sieht so aus, als würden keine neuen Dokumente in den Volltext aufgenommen werden.	100

4 Problemanalyse und Behebung

Falls es zu einem harten Abbruch eines Übersetzungsmoduls oder zu einem hängenden Übersetzungsprozess kommt, kann es vorkommen, dass das Stop-Handshake in einem undefinierten Zustand zurückbleibt. In diesem Fall kann man am einfachsten den OCR Service stoppen. Anschließend kontrolliert man im TaskManager ob alle Übersetzungsprozesse beendet sind. Dieser Vorgang kann bei umfangreichen Dokumenten viele Sekunden oder auch mehrere Minuten dauern.

Anhand der Anzahl der Einzelseiten im Temp-Verzeichnis des Prozesses und der bereits vorhandenen Layout Dateien kann man den Fortschritt dieser Aktion überwachen. Wenn es hier längere Zeit keine Veränderungen mehr gibt, dann ist zu vermuten, dass der OCR Prozess hängt. In diesem Fall kann man diesen Prozess über den TaskManager zwangsweise beenden und die Dateien im Temp-Verzeichnis für eine weitere Analyse sichern und dort manuell löschen. Erst wenn alle Übersetzungsprozesse beendet wurden (d.h. sie werden im TaskManager nicht mehr aufgeführt), können

Sie das Service-Modul neu starten. Mit einer Zeitverzögerung von jeweils 10 bis 15 Sekunden werden die Übersetzungsprozesse dann neu angelegt.

Das Übersetzungsmodul fängt Schutzverletzungen aus dem OCR Prozess ab (z.B. durch defekte TIFF-Dokumente erzeugt) und entfernt diese Datei aus dem Analyseprozess. Je nach Prozesszustand kann es dabei aber vorkommen, dass im Temp-Verzeichnis Dateien zurückbleiben. Aus diesem Grund sollte man hin und wieder eine Kontrolle der Temp-Verzeichnisse auf ältere Dateien vornehmen. Da dort nur aktuelle Arbeitsdateien liegen dürfen, können Sie bei allen Dateien die älter als ein Tag sind, davon ausgehen, dass es sich um solche Überbleibsel handelt. Diese können Sie entfernen.

Defekte Dokumente werden in das Broken-Verzeichnis verschoben. Hier können Sie analysieren, warum der OCR Vorgang abgebrochen wurde:

a) Ungültiges Tiff-Format, z.B. LZW Kompression
b) Ungültige Tiff-Datei, z.B. unvollständig oder beschädigt
c) ...

5 Beispiele für eine manuelle Konfiguration

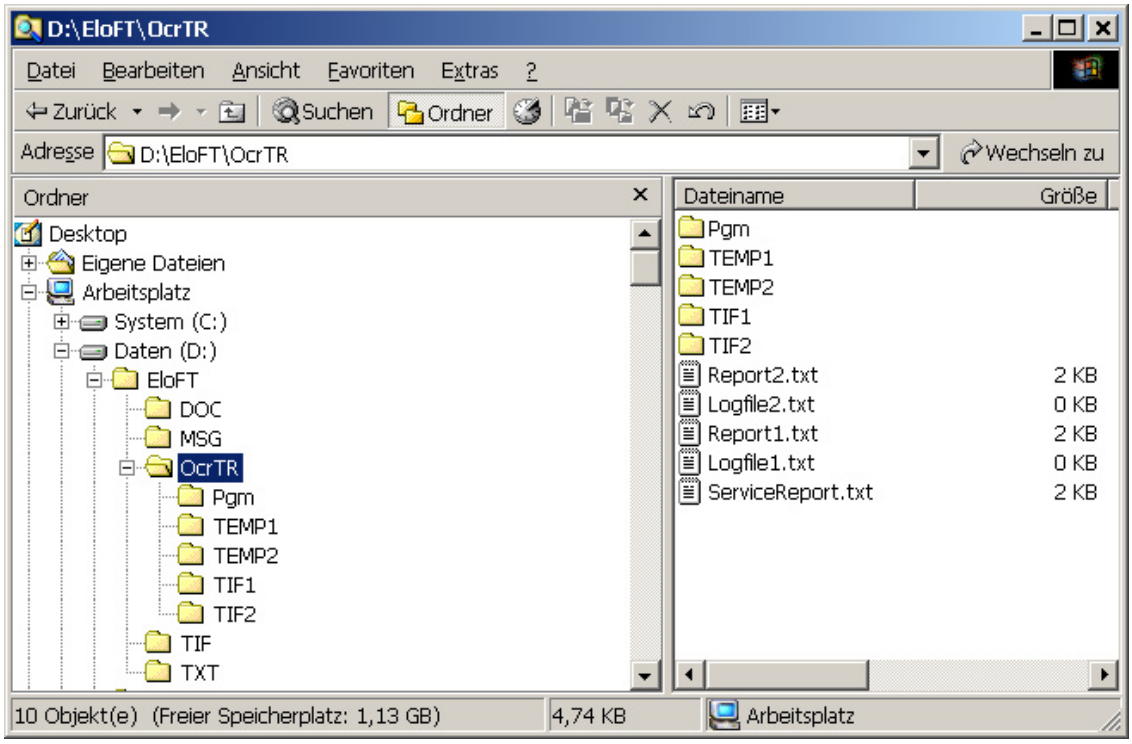
Normalerweise werden die Verzeichnisstruktur und die Registry-Einträge vom Setupprogramm angelegt. In besonderen Fällen kann es aber notwendig sein von den Standardvorgaben abzuweichen. In diesem Fall können Sie die Beschreibungen in diesem Kapitel verwenden.

Im Folgenden finden Sie zwei Reg-Dateien welche eine Beispielkonfiguration für das Service-Modul und das Übersetzungsmodul enthalten. Diese können Sie sich herauskopieren und in einem Texteditor an Ihre spezielle Konfiguration anpassen und anschließend in die Registry importieren.

Dabei wird von folgender Struktur ausgegangen:

1. Der Volltextdienst legt seine Tiff-Dateien im Verzeichnis D:\EloFT\TIF ab.
2. Der Volltextdienst erwartet seine Textdateien im Verzeichnis D:\EloFT\TXT.
3. Für die Übersetzungsmodule gibt es ein Sammelverzeichnis D:\EloFT\OcrTr, dort werden die Report- und die Log-Dateien abgelegt.
4. Unterhalb des OcrTr Verzeichnisses gibt es für jedes Übersetzungsmodul ein TIF* Verzeichnis, also TIF1, TIF2... (nur beim Betrieb mit mehreren OCR Threads)
5. Unterhalb des OctTr Verzeichnisses gibt es für jedes Übersetzungsmodul ein TEMP* Verzeichnis, also TEMP1, TEMP2...
6. Unterhalb des OcrTr Verzeichnisses gibt es für jedes Übersetzungsmodul ein Broken* Verzeichnis, also Broken1, Broken2...
7. Für die Programmdateien gibt es ein Verzeichnis D:\EloFT\OcrTr\Pgm

Diese Verzeichnisstrukturen müssen vor der Konfiguration existieren bzw. angelegt werden. Die genauen Pfade richten sich dabei nach der speziellen Konfiguration vor Ort.



5.1 Reg-Datei für das Service-Modul mit einem OCR Thread

```
[HKEY_LOCAL_MACHINE\SOFTWARE\ELO Digital\FullText]
"LowPriority"=dword:00000000
"OCRThreads"=dword:00000001
"ReloadInterval"=dword:00015180
"ServiceReport"="D:\\EloFT\\OcrTr\\SReport.txt"
```

5.2 Reg-Datei für das Service-Modul mit mehreren OCR Threads

Windows Registry Editor Version 5.00

```
[HKEY_LOCAL_MACHINE\SOFTWARE\ELO Digital\FullText]
"LowPriority"=dword:00000000
"OCRThreads"=dword:00000002
"SplitInput"="D:\\EloFT\\TIF\\"
"ReloadInterval"=dword:00015180
"ServiceReport"="D:\\EloFT\\OcrTr\\SReport.txt"
```

5.3 Reg-Datei für das Übersetzungsmodul mit einem OCR Thread

Windows Registry Editor Version 5.00:

```
[HKEY_LOCAL_MACHINE\SOFTWARE\ELO Digital\FullText.1]

"FineReader"="C:\\Programme\\ABBYY FineReader Engine 7.0\\Bin\\FREngine.dll"

"InputDir"="D:\\EloFT\\ TIF\\"

"OutputDir"="D:\\EloFT\\TXT\\"

"TempDir"="D:\\EloFT\\OcrTr\TEMP1\\"

"Languages"="German,French,English"

"Stop"=dword:00000000

"LogFile"="D:\\EloFT\\OcrTr\\Logfile1.txt"

"Pages"=dword:00000000

"Report"="D:\\EloFT\\OcrTr\\Report1.txt"

"Broken"="D:\\EloFT\\ Broken1\\"
```

5.4 Reg-Datei für das Übersetzungsmodul mit mehreren OCR Threads

5.4.1 Für den OCR Thread 1:

Windows Registry Editor Version 5.00:

```
[HKEY_LOCAL_MACHINE\SOFTWARE\ELO Digital\FullText.1]

"FineReader"="C:\\Programme\\ABBYY FineReader Engine 7.0\\Bin\\FREngine.dll"

"InputDir"="D:\\EloFT\\OcrTr\\TIF1\\"

"OutputDir"="D:\\EloFT\\TXT\\"

"TempDir"="D:\\EloFT\\OcrTr\TEMP1\\"

"Languages"="German,French,English"

"Stop"=dword:00000000

"LogFile"="D:\\EloFT\\OcrTr\\Logfile1.txt"

"Pages"=dword:00000000

"Report"="D:\\EloFT\\OcrTr\\Report1.txt"

"Broken"="D:\\EloFT\\ Broken1\\"
```

5.4.2 Für den OCR Thread 2:

Windows Registry Editor Version 5.00:

```
[HKEY_LOCAL_MACHINE\SOFTWARE\ELO Digital\FullText.2]

"FineReader"="C:\\Programme\\ABBYY FineReader Engine 7.0\\Bin\\FREngine.dll"

"InputDir"="D:\\EloFT\\OcrTr\\TIF2\\"

"OutputDir"="D:\\EloFT\\TXT\\"

"TempDir"="D:\\EloFT\\OcrTr\TEMP2\\"

"Languages"="German,French,English"

"Stop"=dword:00000000

"LogFile"="D:\\EloFT\\OcrTr\\Logfile2.txt"

"Pages"=dword:00000000

"Report"="D:\\EloFT\\OcrTr\\Report2.txt"

"Broken"="D:\\EloFT\\ Broken2\\"
```

6 Mögliche Systemeinschränkungen

6.1 Erkennungsqualität

Die Erkennungsqualität beim OCR Prozess hängt von einer Vielzahl von Parametern ab, die weder der Hersteller der OCR Software noch der Kunde zu 100% bestimmen kann. Deshalb kann es vorkommen, dass einzelne Dokumente mitunter nur eine schlechte Erkennungsrate aufweisen können. Das kann an einer sehr schlechten Scanqualität liegen, aber auch an bestimmten Schriftarten, die für den OCR Prozess nur schlecht geeignet sind.

Folgende Punkte sollte beachtet werden um eine möglichst gute Erkennungsrate zu gewährleisten:

- Die Scanauflösung sollte mindestens 300 dpi betragen.
- Das Dokument sollte möglichst gerade eingescannt werden.
- Die Schwarzschwelle sollte so eingestellt werden, dass die Buchstaben weder auseinander fallen noch dass sie zusammen laufen (z.B. dass ein kleines e im oberen Bereich komplett Schwarz wird).
- Der Scanmodus sollte auf „Line Art" stehen, d.h. es wird ein reines Schwarz/Weiß Bild erzeugt. Dithering führt zu stark ausgefransten Buchstaben, welche die Erkennungsrate deutlich herabsetzen können.
- Reine Textdokumente können wesentlich schneller analysiert werden als Dokumente mit umfangreichen Bildinhalten.

6.2 Dokumente in der „Freien Bearbeitung"

Normalerweise unterliegen ELO Dokumente der Versionsverwaltung. Das führt dazu, dass jede neue Version eine neue Dokumenten-Id erhält, so dass es niemals zu Kollisionen unterschiedlicher Dokumente kommen kann. Bei Dokumenten in der freien Bearbeitung ist das nicht der Fall. Hier wird die aktuelle Dokumentendatei jeweils überschrieben, die Dokumenten-Id bleibt unverändert erhalten.

Wenn so eine Dokument in den Volltext aufgenommen wird und mehrfach schnell hintereinander bearbeitet wird, kann es bei der OCR Analyse mit mehreren Übersetzungsprozessen dazu kommen, dass die Analyse der älteren Version später fertig wird als die Analyse der aktuellen Version. In diesem Fall kann sich eine Volltextsuche nur auf die Daten der älteren Version stützten, da diese nach der aktuellen Version in die Datenbank übernommen wurden. Da wir aber von der freien Bearbeitung ohnehin dringend abraten (aufgrund der fehlenden Versionskontrolle), sollte dieser Fall in der Praxis nicht auftreten.

6.3 Unbekannte Tiff-Formate

Die TIFF-Spezifikation sieht eine Vielzahl von Dokumentenformaten und Kompressionsalgorithmen vor. Zudem sind hier firmenspezifische Erweiterungen eingeplant

(wird z.B. für die Wang-Annotations im Microsoft Viewer verwendet), die nicht immer dokumentiert und zum Teil patentgeschützt sind. Aus diesem Grund kann kein Hersteller garantieren, dass **alle** TIFF-Formate erkannt werden. Wichtig ist, dass die üblichen Formate erkannt werden. Bei Scan-Dokumenten sind dies im Normalfall Dokumente mit der FAX-G4 Kompression.

Falls Sie in Ihrem Datenbestand Tiff-Formate haben, die nicht erkannt werden, dann senden Sie uns bitte ein Beispiel zu. Wir können dann abklären, ob eine entsprechende Konvertierung oder Erweiterung eingefügt werden kann. Im Wesentlichen sind wir hier aber auf die Formate begrenzt, die der Abbyy FineReader einlesen kann.

6.4 Fehler bei der Erkennung

Während des OCR Prozesses werden zahlreiche temporäre Dateien erstellt. Sie werden im Normalfall am Ende des Vorgangs wieder gelöscht. Bei ungewöhnlichen Fehlersituationen kann es vorkommen, dass einzelne Dateien im Temp-Bereich des Übersetzungsmoduls zurückbleiben. Diese führen nicht zu Fehlfunktionen im nachfolgenden OCR Prozess, belegen aber unnötig Speicherplatz. Sie sollten deshalb hin und wieder im Rahmen der normalen Systemwartung entfernt werden. Beachten Sie dabei bitte, dass diese Dateien nicht gelöscht werden konnten, weil sie durch einen Prozess noch gesperrt waren. Im Normalfall wird diese Sperre es nach dem nächsten Reload Zeitpunkt freigegeben. Falls das Löschen also mit einem Sperrhinweis fehlschlägt, nehmen Sie diese Datei aus der Löschliste heraus und entfernen diese erst nach dem nächsten Reload.

6.5 Geschwindigkeit

Ein OCR Vorgang ist eine umfangreiche Aufgabe, die eine erhebliche Prozessorlast erzeugen kann. Falls Sie viele TIFF-Dokumente in den Volltext übernehmen müssen, sollten Sie deshalb eine eigene Maschine für den OCR Textreader vorsehen. Ein Testlauf mit einer größeren Anzahl von normalen Bürodokumenten auf einer Dual Prozessor Maschine hat eine Erkennungsrate von ca. 30.000 Seiten pro Tag gebracht.

Beachten Sie aber bitte, dass diese Zahl in hohem Maße auch vom Inhalt der Dokumente abhängt. Als Faustregel kann man sagen, dass viele eingebundene Bilder die Rate erheblich senken können und maschinell erzeugte Dokumente (die besonders sauber und kantenscharf sind) schneller ausgewertet werden können. Falls Sie für ein Projekt konkrete Zahlen benötigen, dann kann das nur über eine Auswahl von repräsentativen Dokumenten des Kunden und einen Testlauf mit diesen Dateien erfolgen. Für diese Analyse führen die Übersetzungsmodule einen Seitenzähler mit, so dass Sie die Zahl der Seiten pro Stunde für diese Testdokumente exakt ermitteln können.

6.6 Fehler im OCR Prozess

Die OCR Analyse ist eine äußerst komplexe Operation. Bei sehr komplexen Dokumenten kann die Seitenanalyse extrem viel Zeit beanspruchen oder sogar ganz hängen bleiben. Für diesen Fall gibt es im Textreader eine Watchdog-Funktion, welche den Prozess optional nach einer einstellbaren Zeit abbricht (30..200 Sekunden sind im Normalfall günstig) und das betroffene Dokument und Seite speichert. Der OCR Dienst erkennt den Abbruch und startet den Textreader neu. Dieser wird dann zuerst das letzte Dokument bearbeiten, dabei aber die kritische Seite auslassen. Innerhalb eines Dokuments werden bis zu zwei nicht analysierbare Seiten akzeptiert. Wenn es mehr sind, wird das komplette Dokument verworfen.