

MIRA

CONSULTING GMBH

WISSENSMANAGEMENT

LANGZEITARCHIVIERUNG

TEXTERKENNUNG

MIRA CR
Release 2.0

COMIRBAulting



MiraCR Release 2.0

Die MIRA [suite] Anwendung besteht aus mehreren Softwaremodulen, die kooperieren um Dokumente aller Art zu analysieren und in ein Datenbanksystem zu speichern.

Eines der Softwaremodule ist MiraCR, verantwortlich für die Texterkennung von Originaldokumenten verschiedenster Formate (reiner Text, Html, Pdf, gescannte Documente, Bilder, MS-Word, MS-Excel, etc..).

Das neue Release von MiraCR (das frühere „efsdigi“) umfasst einige neue Anwendungsmöglichkeiten und Verbesserungen. Die Auslieferung des Moduls erfolgt in den nächsten Wochen, nachdem die Tests auf unseren Servern beendet sind.

Neue Features :

- Neuer Name "MiraCr anstatt dem früheren "efsdigi"

- Die Parameterdatei für dieses Modul heißt MIRA.INI. Es ist der erste Ansatz in der Re-Organisation der MIRA [suite] um die Wartung und Pflege zu vereinfachen.

- MiraCR kann jetzt auch Volltext und Meta-Informationen aus OpenOffice Dokumenten (Version 1 und 2) extrahieren. OpenOffice ist ein Open Source Paket was von der Funktionalität der MS Office Suite gleicht, welches aber unter verschiedenen Betriebssystemen (MS-Windows, Solaris, Linux, HUX...) verfügbar ist.

- MiraCR kann die Inhalte von ZIP-Dateien bearbeiten. Mit der richtigen Einstellung kann MiraCR aus den einzelnen Komponenten von ZIP-Dateien den Text extrahieren und den Inhalt sowie die ZIP-Datei selbst auf verschiedenste Arten weiterverarbeiten. Das neue Feature wird durch ein neues Tag in MIRA.INI <ZipProcess>-Datei verwaltet. Die Grundeinstellung des Prozesses ist „blob“. Die ZIP-Datei wird als eine Datei ohne weitere Verarbeitung behandelt. Es erfolgt keine Texterkennung.

Weitere Neuerungen :

Einige Änderungen wurden innerhalb des Programms vorgenommen, um die Stabilität bezüglich Netzwerkproblemen zu verbessern. Die meisten Anpassungen sind technischer Natur und werden deshalb nicht alle ausführlich beschrieben.

→ In der früheren Version wurde eine „document source“ ausgeschlossen, wenn der MiraCR Prozess wegen Netzwerkproblemen nicht angesprochen werden konnte. Erst nach dem Neustart des MiraCR Services wurde die Source wieder eingebunden, wenn sie wieder zur Verfügung stand. In der neuen Version wird MiraCR die Quelle nur kurzfristig ausschließen und nach vordefinierter Zeit (Voreinstellung 15 min) wieder versuchen die Quelle anzusprechen.

→ Dieser Parameter kann in der MIRA.INI – Datei <RetryAfter> - Parameter eingestellt werden.

Details des ZIP Filing – Prozesses

Ein neuer <ZipProcess> Parameter wurde in der MIRA.INI Parameterdatei in der <Source> Sekktion eingeführt. Dieser Parameter bietet vier verschiedene Einstellungsmöglichkeiten zum Verarbeiten der ZIP-Dateien an:

→ **blob** : definiert, dass die ZIP – Datei als “binary file” verarbeitet werden soll. Die Datei wird nicht durch den Prozess gelesen sondern an das nächste Modul weitergereicht. Das heißt die Handhabung ist dieselbe wie es im Vorgängermodul behandelt wurde. Das ist die “default” - Einstellung, was bedeutet, dass die Datei gleich wie bisher im „efsdigi“ behandelt wird

→ **extract** : Die original ZIP-Datei wird als “blob” verarbeitet. Zusätzlich werden die einzelnen Komponenten mit der Texterkennung und Meta-Daten Gewinnung bearbeitet. Jede einzelne Datei des ZIP-Files wird als eigenes Dokument verarbeitet und im Filing Volume abgelegt.

→ **dig** : Die original ZIP-Datei wird als “blob” verarbeitet. Zusätzlich werden die einzelnen Komponenten mit der Texterkennung und Meta-Daten Gewinnung bearbeitet. Jede einzelne Datei des ZIP-Files wird als eigenes Dokument verarbeitet. Ein logischer Link wird in der Datenbank gehalten und verweist auf die ursprüngliche ZIP-Datei. Aber im Gegensatz zur extract Methode werden die Komponenten nicht im Filing Volume abgelegt sondern nach der Verarbeitung gelöscht. Die Texterkennungsergebnisse werden je in einem Datensatz gespeichert und mit einem Link zur ZIP-Datei in der Datenbank abgelegt.

→ **replace** : Die original ZIP-Datei wird verworfen, nur die einzelnen Komponenten werden im Filing Volume abgelegt und mit einem gemeinsamen Link versehen, damit die ZIP-Datei logisch wieder hergestellt werden kann.

Wir hoffen, dass diese neuen Optionen alle Wünsche die bezüglich ZIP-Datei Verarbeitung auftreten damit erfüllt werden können. Aber wir denken, dass einer unserer Nutzer sicher noch eine weitere Variante findet. MiraCr ist so designet, dass es relativ einfach ist weiter Features anzubinden.

Die neue MIRA.INI Parameterdatei

Wie die bisherigen Nutzer der Mira [suite] wissen, läuft das Programm sehr zuverlässig. Auch wenn es konfiguriert ist, sind Anpassungen, Ergänzungen zur Konfiguration einfach und flexible möglich.

Diese Flexibilität hat den Preis einer gewissen Komplexizität bei der Installation und etwaiger Problembhebung wenn etwas nicht rund läuft.

Einer dieser Gründe ist, dass MIRA selbst aus verschiedenen Modulen besteht die auf verschiedenen Hosts/Server, komplexen Netzwerkstrukturen und verschiedenen Betriebssystemen installiert sein können.

Unser mittelfristiges Ziel ist, eine Art "Kontrollzentrum" zur Verfügung zu stellen um dem Administrator bzw. installierendem Techniker die Installation, Updates und Verwaltung mit einem browserbaxierenden Werkzeug zu ermöglichen. Als erster Schritt in diese Richtung haben wir beschlossen alle Parameter in einem zentralen INI-File für alle Module zusammen zuführen. Das ist das neue MIRA.INI File, welches in Zukunft die verschiedenen "per-module" INI-Files (EFSDIGI.INI, EFSLOADX.INI, etc..) ersetzen wird. Das neue MiraCR ist das erste Modul was das neue zentrale MIRA.INI nutzt. Die anderen Module werden demnächst folgen.

Das Format der MIRA.INI Datei sieht folgendermaßen aus:

- eine startende <General> Sektion, die die gültigen Parameter für alle Module enthält.

- dann folgt eine Sektion pro Modul (<MIRACR> für das MiraCR Modul, <MIRAFILER> für das Filer Module, etc..

Es soll, auch wenn mehrere Hosts involviert sind, nur eine MIRA.INI Datei zur Verwaltung geben. Der Masterhost wird dann auch ein Interface zur Verwaltung, Kontrolle und Konfiguration aller Module auf allen Servern zur Verfügung stellen.

Alle Module werden ihre Konfiguration beim Masterhost anfordern.

MIRA

CONSULTING GMBH

MIRA Consulting GmbH
Bruckwiesenstraße 1
D - 72336 Balingen

Telefon: 0 74 33/90 72 31.0
Telefax: 0 74 33/90 72 31.18

E-Mail: info@mira-consulting.net
www.mira-consulting.net

Vorsprung durch
Wissen.

