

MIRA

CONSULTING GMBH

WISSENSMANAGEMENT

LANGZEITARCHIVIERUNG

TEXTERKENNUNG

MIRA CR
Release 2.0

COMIRBAulting



MiraCR Release 2.0

The MIRA application consist of a series of software modules, co-operating to process original business documents, analyse them, and save them in a database system in a way which allows their easy and efficient retrieval when needed.

One of the software modules of the MIRA Suite is MiraCR, the module responsible for extracting text and meta-information from a variety of original document formats (plain text, html, pdf, scanned document images, MS-Word, MS-Excel, etc..).

Until now, this module was a program called "efsdigi". We are pleased to announce a new version of this module, which includes a couple of new features and improvements. The new version will be available in the coming weeks, as soon as the period of intensive testing on our own servers is finished.

New features :

- the new program is called "MiraCr (instead of efsdigi).

- the parameter file for this module is called MIRA.INI. This is the first step in a new organisation of the MIRA Suite, aimed at simplifying maintenance and support. More information follows.

- MiraCR can now extract text and meta-information out of OpenOffice documents (versions 1 and 2), in addition to the formats it already handled. OpenOffice is an Open Source package functionally equivalent to the Microsoft Office Suite, and which works in several computer environments (MS-Windows, Solaris, Linux, HPUX,..).

- MiraCR is now capable of handling ZIP files as something other than “blobs”. With the appropriate setup, MiraCR can now optionally extract the components of ZIP files, and automatically process the original ZIP file and the individual components in a variety of ways. This new feature is enabled by a new parameter <ZipProcess> in the MIRA.INI file, and is described more in detail below. The default value of <ZipProcess> is “blob”, which means that by default it will do as before : the ZIP file will be filed as an original attachment, but no text or meta-information are extracted from it.
-

Other changes :

Several changes have been made inside the program, mainly regarding better stability when faced with network-related problems, and to improve the automatic recovery in case of incident. Most of these changes are fairly technical and are not detailed here, but one of the changes has a practical aspect that some users may be aware of :

- in the previous versions of the MiraCR module, when a “document source” became inaccessible because of a network error, MiraCR automatically eliminated that source from further processing, until MiraCR was restarted. In the new version, MiraCR will only temporarily disable the bad source, and will automatically retry it after a period of time. By default, this period is set to 15 minutes, but it can be adjusted via a new <RetryAfter> parameter in the source description (in MIRA.INI).

Details of ZIP file processing

A new <ZipProcess> parameter has been introduced in the MIRA.INI parameter file's <Source> sections. This parameter can take one of 4 values, each representing a way in which MiraCR should process any ZIP-file attachments arriving for processing in that source directory:

- **blob** : indicates that MiraCR should process ZIP files as “binary files” : file them as original, but do not “open” the file and scan it's content. This the default, and is the same as what happens until now with ZIP files (efsdigi program). In other words, if you do not add this parameter to a <source> section, nothing changes.

- **extract** : the original ZIP file is filed as a “blob” (like in the default), but in addition each ZIP file component is extracted, and processed individually as if it was a separate attachment. That means that each component is extracted from the ZIP file, and processed for text and meta-data extraction (depending on it's format), then filed as an original document. A logical link is kept in the Star database, to indicate that this was a component of the original ZIP file.

- **dig** : the original ZIP file is filed as a blob (like in the default), but in addition each ZIP file component is extracted, and processed individually as if it were a separate attachment : the text and meta-information are extracted, and passed to the data base. But, contrary to the extract method above, the extracted component is discarded after text extraction and not filed as an original document. In other words, you will get the text and indexing of the document in the database, but no link to the original.

- **replace** : the original ZIP file is discarded, but it is replaced by it's components, which are processed (and filed) individually as originals. This would have the same result as if there never was an original ZIP file, but instead there were the extracted components.

We sincerely hope that the options above cover what most users might ever want to do with ZIP files and their components. But we expect that someone will find a variation that is not yet covered, and MiraCR has been designed in order to make it fairly easy to adopt such additional needs if required.

the new MIRA.INI parameter file

As the current users of the MIRA Suite know, once it is installed and configured, it is very reliable, very easy to use and very flexible. That is good, and that is how it was designed to be. But this ease and flexibility for the user comes at the price of a certain complexity in the initial setup, and in troubleshooting when something goes wrong. One of the reasons for this is that MIRA consists of several modules, each of which can be installed on a different host computer, in various kinds of network environment and under various operating systems.

One of our development goals in the medium-term is to provide a central “control center” where the system administrator and the installation technician can install, update, and control the workings of the whole MIRA suite through a single interface (probably a web page).

As a step in that direction, we have decided to “centralise” in the future all the parameters of the various MIRA Suite modules in one central location. This is the new MIRA.INI file, which in the future will replace the various “per-module” INI-files (EFSDIGI.INI, EFSLO-ADX.INI, etc..).

The new MiraCR module is the first to use this new central MIRA.INI file. Other modules will be rewritten shortly to do the same.

The format of this MIRA.INI file is as follows :

- an initial <General> section, which will contain parameters valid for all the modules

- then one section per module (e.g. <MIRACR> for the MiraCR module, <MIRAFILER> for the filer module, etc..)

The idea for now is that if there are several host servers involved, one single MIRA.INI file can be maintained, and can be simply copied to all host servers for all MIRA modules. In the future, there will probably be only one MIRA.INI file on one single server; this single server will then host a new control module of the MIRA Suite, from which all other modules will request their individual configuration. This control module will also provide an interface to the system administrator, to enquire about the status of the various MIRA modules, make changes to their configuration, and to control them.

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.

