ELO Textreader

[Stand: 02.10.2019 | Programmversion: 12.00.005.001]

Das Programm ELO Textreader ist ein Servlet, das aus verschiedenen Dokumententypen Text extrahieren und in einer Textdatei speichern kann.

Inhalt

1 Allgemeines 4
1.1 Aufgabe des Programms 4
1.2 Versionsinformationen 4
1.3 Dateiformate 5
1.4 Voraussetzungen 7
1.5 Programminstallation 8
1.5.1 Update der Textreader-Konfiguration auf die aktuelle Version (ab Textreader 10) 8
1.5.2 Neuinstallation des Textreaders 9
1.6 Logging-Ausgaben 10
2 Konfiguration über ‚config.xml' (bis Textreader 9) 12
2.1 Aufbau der ‚config.xml' (bis Textreader 9) 12
2.2 Beispiele für Einträge in der ‚config.xml' 20
2.2.1 Erläuterungen zur Konfiguration der Dokumentengröße 20
2.2.2 Erläuterungen zur Konfiguration von PDF-Dokumenten 20
2.2.3 Erläuterungen zur Konfiguration von MMF-Dokumenten 21
2.2.4 Erläuterungen zur Konfiguration von WMF-Dokumenten 22
2.2.5 Erläuterungen zur Konfiguration von TIF-Dokumenten 22
2.2.6 Erläuterungen zur Konfiguration von 7-Zip-Dokumenten 22
2.2.7 Erläuterungen zur Konfiguration von minimaler Breite und Höhe 22
2.2.8 Erläuterungen zur Konfiguration des Verhaltens im Fehlerfall 22
3 Konfiguration über die ELO Administration Console (ab Textreader 10, beschrieben für Version 10.01.000) 24
3.1 Aufbau der ‚config.xml' ab Textreader 10 24

3.2 Bereich ‚Rückgabeverzeichnis' ... 24
3.3 Bereich ‚Ausgabeverzeichnisse und typspezifische Konfiguration' ... 26
3.3.1 Identisch zu konfigurierende Dateitypen ... 27
3.3.2 Option ‚Extern konvertieren' ... 28
3.3.3 Optionen für den Dateityp ‚pdf' ... 29
3.3.4 Optionen für den Dateityp ‚tif(f)' ... 29
3.3.5 Konvertierung von Textdateien ... 30
3.3.6 Ein weiteres Verzeichnis hinzufügen ... 30
3.3.7 Ausgabeverzeichnisse entfernen ... 30
3.4 Bereich ‚Textreader Schwellenwerte' ... 31
3.4.1 Option ‚Mindestbreite oder Mindesthöhe' ... 31
3.4.2 Option ‚Maximale Fehlerzahl' ... 31
3.4.3 Option ‚Maximal exportierte Dateien pro Minute' ... 31
3.4.4 Option ‚Maximal importierte Dateien pro Minute' ... 32
3.4.5 Option ‚Maximale Fehlerzahl pro Tag' ... 32
3.4.6 Option ‚Maximale Dateigröße in MB' ... 32
3.4.7 Option ‚Zeitspanne pro Seite im OCR-Service' ... 32
3.4.8 Option ‚Zeitspanne pro OCR-Dokument im Textreader' ... 33
3.4.9 Option ‚Mindestspeicherplatz in Prozent' ... 33
3.4.10 Option ‚Mindestspeicherplatz absolut' ... 34
3.4.11 Option ‚Maximale Anzahl Dateien im Ordner' (bis Textreader-Version 11.01) 34
3.4.12 Option ‚Maximale Anzahl Zeichen pro Zeile in MSG-Dateien' ... 35
3.4.13 Option ‚Maximale Dateigröße einer Volltextdatei' ... 35
3.5 Bereich ‚Textreader weitere Konfiguration' ... 35
3.5.1 Option ‚Erweiterter Trace' ... 35
3.5.2 Option ‚Default-Ordner' ... 36
3.5.3 Option ‚Freigabezeiten für Export und Import' ... 36
3.5.4 Option ‚Allgemeines Verhalten im Fehlerfall' ... 36
3.6 Bereich ‚OCR-Optionen' ... 38
3.6.1 Option ‚OCR-Lesemodus' ... 38
3.6.2 Option ‚OCR-Erkennungsmodus ausführlich' ... 38

3.6.3 Option ‚OCR-Sprachen' ... 38
3.6.4 Option ‚Anzahl der benutzten OCR-Verbindungen' ... 39
4 Default-Werte der Konfigurationsoptionen nach Neuinstallation der aktuellen Textreader-Version (ab Version 10) ... 41
4.1.1 Abschnitt ‚Ausgabeverzeichnisse und typspezifische Konfiguration' ... 41
4.1.2 Abschnitt ‚Textreader Schwellenwerte' ... 41
4.1.3 Abschnitt ‚Textreader weitere Konfiguration' ... 42
4.1.4 Abschnitt ‚OCR-Optionen' ... 42
4.1.5 Abschnitt ‚Verhalten bei Fehlern' ... 42
5 Anpassung von Default-Werten in der Datenbank (ab Version 10) ... 43
5.1 Abschnitt ‚Ausgabeverzeichnisse und typspezifische Konfiguration' ... 43
5.2 Abschnitt ‚Schwellenwerte' ... 44
5.3 Abschnitt ‚Textreader weitere Konfiguration' ... 44
5.4 Übersicht für Optionen in der Datenbank ... 46

1 Allgemeines

Information: Dieses Handbuch ist gültig für alle Textreader-Versionen ab Textreader 9.

1.1 Aufgabe des Programms

Das Programm ELO Textreader ist ein Servlet, das aus verschiedenen Dokumententypen Text extrahieren und in einer Textdatei speichern kann. Die zu extrahierenden Dokumente können sich auch in einem ZIP- oder 7-Zip-Archiv oder auch als Anhang in einer E-Mail (MSG, EML) befinden. Je nach Dateiformat wird zusätzliche Funktionalität unterstützt. Aus Dateien im PDF-Format können auch Bilddateien extrahiert werden. Das ab Textreader-Version 9.17.040.000 integrierte ehemalige ELO Fulltext Modul (ELOft) wird nun schon ab einer Indexserver-Version >= 9.00.060 gestartet (bisher: erst ab Indexserver-Version >= 10). Auf der Statusseite des Textreaders können Exporter und Importer sowie die drei Threads für PDF-, OCR- und sonstige Konverter gezielt gestoppt und gestartet werden.

Information: Ab Textreader-Version 9.17.040.000 ist der ELOft in den Textreader integriert und wird in Kombination mit einer Indexserver-Version >= 9.00.060 gestartet, der ELOft muss dann deinstalliert bzw. deaktiviert werden.

Information: Ab Textreader-Version 12 können Exporter, Importer und Konverter auf der Statusseite gezielt gestoppt und gestartet werden.

1.2 Versionsinformationen

Der Textreader ist sowohl mit Indexserver 9 (ab Version 9.00.060) als auch mit Indexserver 10, 11 und 12 lauffähig. Ab Version 11 ist die Versionsnummer wieder umgestellt worden, entsprechend der Systematik <Hauptrelease>.<Nebenrelease>.<Patch-Nummer>.<Build-Nummer>, hier also 12.00.000.003. Mit einer Indexserver-Version kleiner gleich 10 stellt sich die Versionsnummer dar wie bisher, nämlich <IX-Versionsnummer>.<Jahr>.<Monat><Hotfix>.<Build-Nummer>; die Indexserver-Versionsnummer wird also der Textreader-Versionsnummer vorangestellt, z. B. 10.16.090.001.

Entsprechend sprechen wir von der Textreader-Version 9 oder Textreader-Version 10 bzw. 11 oder 12, die sich in ihrer Funktionalität unterscheiden:

- Ab ELO Textreader-Version 10 erfolgt die Konfiguration des neuen, integrierten Textreaders in der ELO Administration Console, nicht mehr in einer *config.xml*.

- Als ELO Textreader-Version 9 erfolgt die Konfiguration der Konverter wie bisher in der *config.xml*, die Konfiguration der Export- und Importprozesse des nun integrierten ELO Fulltext Moduls in der Datenbanktabelle *eloftopt* (wie bisher auch; Dokumentation dazu finden Sie im Server-Handbuch Version 9, Kapitel „ELO Fulltext“).
- Ab ELO Textreader-Version 10 wird standardmäßig zur Konvertierung von PDF-Dateien der Apache-PDFBox-Konverter benutzt (In der ELO Administration Console kann die OCR zur Konvertierung von PDF-Dateien zurzeit nicht mehr eingestellt werden. Grund: ein Fehler im Abbyy Finereader Version 11).
- Ab ELO Textreader-Version 12 können in ELO verschlüsselte Dokumente konvertiert werden.



Information für ältere Textreader-Versionen: Da die PDFBox in Kombination mit älteren Textreader-Versionen zum Absturz des Tomcat führen konnte, sollte bei der Installation von Textreadern mit Versionsnummer kleiner als 9.03.006.001 der ELO OCR-Service 9 mit Abbyy Finereader Version 10 zur Konvertierung von PDF-Dateien genutzt werden. Wobei die PDFBox weiter zu Verfügung stünde, z. B. bei schlechter Qualität der OCR-Erkennung, dann aber der Textreader in einem separaten Tomcat installiert werden sollte (also in einem eigenen Adressraum), damit beim Absturz der PDFBox nicht alle weiteren Servlets zum Absturz gebracht würden.

1.3 Dateiformate

Die zu extrahierenden Dokumente werden aus dem ELO Archiv in Verzeichnisse exportiert, deren Namen dem jeweiligen Dateityp entsprechen, und so dem Konverter zur Verfügung gestellt. Nach der Konvertierung werden die resultierenden Volltext-Dateien wieder in das ELO Archiv importiert.

Folgende Formate werden unterstützt:

- AI – Adobe Illustrator
- BMP – Windows Bitmap
- CSV – Comma-separated values
- DOC – Microsoft Word Textverarbeitungsprogramm
- DOCX – Microsoft Word Textverarbeitungsprogramm ab Version Word 2007
- DOT – Microsoft Word Textverarbeitungsprogramm
- DXL – Domino XML (exportierte IBM Notes-Dokumente)
- EML – Microsoft Outlook Express E-Mail Format
- GIF – Graphics Interchange Format
- GZ – gzip (“GNU zip”)

- HTM, HTML – Hypertext Markup Language
- JPG, JPEG – (Joint Photographic Experts Group)
- LOG – Logdateien verschiedener Programme
- MHT – Microsoft Internet Explorer Web Archiv Format
- MMF – ELO COLD (setzt WMF Konverter voraus)
- MSG – Microsoft Outlook E-Mail Format
- ODT – Offenes Dokumentformat für Büroanwendungen (OpenDocument)
- PDF – Portable Document Format
- PNG – Portable Network Graphics
- PPS – Microsoft PowerPoint Show (Dateien müssen in den PPT-Ordner geschoben werden. Legen Sie keinen PPS-Ordner an.)
- PPT – Microsoft PowerPoint Präsentationsprogramm
- PPTX – Microsoft PowerPoint Präsentationsprogramm ab Version PowerPoint 2007
- RTF – Microsoft Rich Text Format
- TIF, TIFF – Tagged Image File Format
- VCF – vCard „elektronische Visitenkarte“
- VSD – Microsoft Visio
- VSDX – Microsoft Visio 2013
- WMF – Windows Metafile Microsoft Grafikformat
- XLA – Microsoft Excel Tabellenkalkulationsprogramm
- XLS – Microsoft Excel Tabellenkalkulationsprogramm
- XLSX – Microsoft Excel Tabellenkalkulationsprogramm ab Version Excel 2007
- XLT – Microsoft Excel Tabellenkalkulationsprogramm (Vorlagen)
- XML – Extensible Markup Language
- ZIP – Ein gängiges offenes Format zur komprimierten Archivierung von Dateien
- 7-ZIP – Ein weiteres gängiges offenes Format zur komprimierten Archivierung von Dateien

Nachfolgende zusätzliche (Header-)Informationen bzw. eingebettete Dokumente werden aus folgenden Dokumentenformaten extrahiert:

- DOCX, XLSX, PPTX – alle eingebetteten Dateien, die dem Textreader bekannt sind
- EML – *Von*, *An*, *CC*, *Betreff* und alle Anhänge
- MSG – *Von*, *An*, *CC*, *Betreff* und alle Anhänge

- PDF – eingebettete Grafiken, JPG, JPEG, PNG, TIF

1.4 Voraussetzungen

Beachten Sie: Diese Textreader-Version benötigt Java 8.

Beachten Sie: Diese Textreader-Version benötigt einen Indexserver mit einer Version >= 9.00.060. Ist der Textreader mit einer Indexserver-Version < 9.00.060 verbunden, muss zusätzlich ein ELOft installiert werden. Der Textreader läuft zwar mit solchen Indexserver-Versionen, räumt aber nach Programmstart $$$-Dateien auf, die temporär vom ELOft produziert werden; Dokumente könnten nicht in den Volltext aufgenommen werden.

Beachten Sie: Diese Textreader-Version benötigt mindestens den ELO OCR-Service Version 10.00.000 und mindestens ELO Indexserver 9.00.060.

Beachten Sie: Ab Textreader-Version 9.02 wird nicht mehr automatisch als OCR-Texttyp *TT_Matrix* gesetzt (dadurch war die Erkennung von Lineprinter-Ausdrucken im Querformat verbessert worden). Stattdessen ist *TT_Matrix* unter *ocr_languages* in der *config.xml* einzutragen, falls solche Dokumente durch die OCR konvertiert werden sollen.

Sie brauchen *TT_Matrix* nicht mehr zu setzen, wenn Sie den Textreader und den ELO OCR Service mit einer Versionsnummer ab 10.00.000 verwenden.

Beachten Sie: Der Textreader benutzt zur Konvertierung von Bilddateien sowie, bis Textreader-Version 9.03.006.000, auch von PDF-Dateien als Default den ELO OCR-Service, wobei ausschließlich der R-Betriebsmodus genutzt wird. Stellen Sie sicher, dass der ELO OCR-Service entsprechend konfiguriert ist. Näheres finden Sie im Handbuch zum ELO OCR-Service.



Beachten Sie: Das ELO Fulltext Modul ELOft muss ab einer Textreader-Version 9.17.040.000 in Kombination mit einer Indexserver-Version ab 9.00.060 deinstalliert bzw. deaktiviert werden.

1.5 Programminstallation

Das Programm ist im ELO Installationsprogramm enthalten.

Bis Version 9 erfolgte die Konfiguration in der *config.xml*. Ab Version 10 wird der Textreader über die ELO Administration Console konfiguriert. Da diese Anleitung aber auch beim Einsatz älterer Versionen noch Verwendung finden soll, wird in dieser Dokumentation auch die Konfiguration mittels *config.xml* beschrieben.

Beachten Sie bei manuellen Updates von früheren Versionen (also **ohne** Aufruf des ELO Installationsprogramms), dass die bisherigen ELO OCR-Programme oder -Dienste deinstalliert bzw. deaktiviert werden müssen.

Bei Updates auf Version 9 beachten Sie die Änderungen in der *config.xml* bzgl. des Dateityps *pdf* und die zusätzlichen Einträge für den OCR-Service und die Konvertierung von Bildern (näheres siehe weiter unten).

Beim Update auf Versionen ab Textreader 10 beachten Sie die Konfiguration in der ELO Administration Console, siehe *1.5.1 Update der Textreader-Konfiguration auf die aktuelle Version (ab Textreader 10)*.

Das manuelle Update einer bestehenden ELO 9 Installation (Textreader-Version <= 9.03.006) durch eine Installation des aktuellen ELO Textreaders und der aktuellen OCR ist beschrieben in dem separaten Dokument *ELO TR 10 für ELO 9 installieren*.

1.5.1 Update der Textreader-Konfiguration auf die aktuelle Version (ab Textreader 10)

Falls ein Update von einer früheren Textreader-Version auf die aktuelle Textreader-Version erfolgen soll, wird durch das Setup-Programm die *config.xml* des Textreaders gelesen und die Optionen in die Datenbanktabelle *eloftopt* eingetragen. Anschließend wird die *config.xml* nach *config.xml.BAK* gesichert und bis auf die Indexserver-Anmeldeinformationen gelöscht. In der *config.xml* verbleiben also lediglich die URL des Indexservers sowie Login-Name und Passwort. Kann keine BAK-Datei angelegt werden, wird die *config.xml* nicht überschrieben.



Beachten Sie: Teilweise werden bei einem Update auch Default-Werte in die Datenbanktabelle *eloftopt* eingetragen. Überprüfen Sie auf jeden Fall nach einem Update die Textreader-Konfiguration (in der ELO Administration Console).

1.5.2 Neuinstallation des Textreaders

Nach einer Neuinstallation durch das Setup-Programm werden nur die Indexserver-Anmeldeinformationen in die *config.xml* geschrieben. In die Tabelle *eloftopt* werden die Pfadnamen der Eingangsverzeichnisse der Konverter, das Ausgabeverzeichnis und verschiedene Default-Werte geschrieben. Diese finden Sie unter *4 Default-Werte der Konfigurationsoptionen nach Neuinstallation der aktuellen Textreader-Version (ab Version 10)*. Sie können diese Optionen mit der ELO Administration Console anpassen, siehe *3 Konfiguration über die ELO Administration Console (ab Textreader 10, beschrieben für Version 10.01.000)*.

1.6 Logging-Ausgaben

Alle Logging-Ausgaben erfolgen bis Textreader-Version 11 in einer Logging-Datei, die in der *log4j.properties*-Datei konfiguriert werden kann.

Ab **Textreader-Version 12** erfolgen die wichtigsten Logging-Ausgaben zusätzlich in dem sogenannten „**Reportlog**". Dort wird nur protokolliert, ob ein Dokument exportiert, importiert, konvertiert oder an einen anderen Konverter oder eine andere Anwendung weitergegeben wurde (insbesondere die OCR). Ebenfalls wird mit Angabe des Grundes protokolliert, wenn ein (Teil-)Dokument nicht konvertiert werden konnte.

Die Protokollierung kann eingeschränkt oder erweitert werden zum Beispiel durch die Konfiguration des Protokollierungs-Levels in der *log4j.properties*-Datei:

- Im **„Error"**-Level wird protokolliert, wenn Dokumente nicht konvertiert werden konnten.
- Im „**Warn**"-Level werden auch sonstige Probleme protokolliert, zum Beispiel unbekannte Dateitypen von Anhängen in MSG- oder EML-Dateien.
- Im „**Info**"-Level wird zusätzlich protokolliert, wenn ein Dokument konvertiert worden ist.
- Im „**Debug**"-Level wird schließlich noch protokolliert, wenn Dokumente exportiert oder importiert werden konnten, oder an eine andere Anwendung übergeben worden sind, zum Beispiel Bilddateien an die OCR.

Hier ein Beispiel einer *log4j.properties*-Datei. Der Logging-Level für den *Reportlog* wurde auf „Info" eingestellt unter „log4j.logger.reportlog=info, reportlog":

```
log4j.rootLogger=info, FI
log4j.logger.reportlog=info, reportlog
log4j.additivity.reportlog=false

# uncomment the following line for debug output:
#log4j.logger.de.elo=debug
#log4j.logger.org.apache.pdfbox=fatal

# output in file:
log4j.appender.FI=org.apache.log4j.DailyRollingFileAppender
log4j.appender.FI.File=e:/temp/log/tr-elo10.log
log4j.appender.FI.DatePattern='.'yyyy-MM-dd'.txt'
log4j.appender.FI.layout=org.apache.log4j.PatternLayout
log4j.appender.FI.layout.ConversionPattern=%d{ABSOLUTE} %t %1x %-5p (%F:%L)
- %m%n
log4j.appender.FI.append=true

log4j.appender.reportlog=org.apache.log4j.RollingFileAppender
log4j.appender.reportlog.File=e:/temp/log/reportlog.log
# DatePattern: one file each day:
log4j.appender.reportlog.DatePattern='.'yyyy-MM-dd'.txt'
```

```
log4j.appender.reportlog.layout=org.apache.log4j.PatternLayout
log4j.appender.reportlog.layout.ConversionPattern=%d{ABSOLUTE} %t %1x %-5p
- %m%n
log4j.appender.reportlog.append=true
```

Der Protokollierungs-Level ist nicht nur in der *log4j.properties*-Datei änderbar, sondern auch auf der Statusseite des Textreaders (im „Edit"-Modus). Gleichzeitig können auf der Statusseite auch neue, leere Log-Dateien angefordert werden.

Logging		
Level	INFO	INFO
Log file	c:/temp/logs/tr-elo12.log	Start new log file
Report log file	c:/temp/logs/reportlog.log	Start new report log file

Abb. 1: Ändern des Protokollierungs-Levels und Erstellen neuer Log-Dateien auf der Statusseite des Textreaders

Der Protokollierungs-Level des *Reportlogs* kann hier aber nur simultan mit dem Protokollierungs-Level des Standard-Loggers umgestellt werden.

Die Ausgabe erfolgt in einem festgelegten Format, erklärt am Beispiel der Datei „00000003.xls" (der Dateiname von exportierten ELO Dokumenten besteht immer aus der Dokumenten-Id in hexadezimaler Form, gefolgt von dem Dateityp):

```
INFO  - HexId >00000003< DocId >3< Filename >00000003.xls< Converted
>C:\ELO11\data\tr-elo110\xls\00000003.xls<
```

Ausgegeben wird sowohl in hexadezimaler als auch dezimaler Form die ELO Dokumenten-Id des zu konvertierenden Dokuments, dann der Dateiname und der Dateipfad, jeweils in spitzen Klammern (zum besseren Filtern der Logging-Datei).

Bei Elementen von sogenannten Container-Dokumenten wird als HexId und DocId immer die Dokumenten-Id des ELO Dokuments ausgegeben. Wenn also z. B. die Datei „0000005E.zip" eine HTML-Datei enthält, wird diese unter dem Namen „0000005E_000.html" extrahiert. Nach der Konvertierung dieser HTML-Datei enthält die Logging-Datei den Eintrag:

```
INFO  - HexId >0000005E< DocId >94< Filename >0000005E_000.html< Converted
>C:\ELO11\data\tr-elo110\htm\0000005E_000.html<
```

2 Konfiguration über ‚config.xml' (bis Textreader 9)

> **Beachten Sie**: Läuft dieser Textreader als Version 9, ist also mit einem ELO Indexserver Version 9 verbunden, wird die Konfiguration wie bisher aus der *config.xml* gelesen.

2.1 Aufbau der ‚config.xml' (bis Textreader 9)

Auszug aus der Datei *config.xml*

```
<comment>parameters for this web application</comment>
<entry key="dirs_delete">C:\ELOenterprise\data\ft elo\delete</entry>
<entry key="dirs_eml">
  C:\ELOenterprise\data\ft-elo\eml|C:\ELOenterprise\data\ft-elo\txt
</entry>
<entry key="dirs_pdf">
  C:\ELOenterprise\data\ft-elo\pdf|C:\ELOenterprise\data\ft-elo\txt|ocr
</entry>
<entry key="dirs_tif">C:\ELOenterprise\data\ft-elo\tif</entry>
<entry key="file_max_size_MB">3</entry>
<entry key="file_max_error">50</entry>
<entry key="convert not possible value">notconv|notconv_sub</entry>
```

Die Textreader-Parameter werden in den <entry>-Tags in der *config.xml* Datei hinterlegt. Ein Parameter besteht aus einem Namen, der im Attribut *key* des <entry>-Tags eingetragen ist, und dem dazugehörigen Wert. Der Wert des Parameters kann aus mehreren Teilen bestehen, die jeweils durch ein | getrennt sind (siehe oben z. B. *dirs_pdf*).

dirs_pdf bedeutet, dass der Textreader Dokumente vom Typ PDF (siehe Übersicht oben) bearbeiten kann.

In *dirs_tif* ist (in diesem Beispiel) nur ein Wert eingetragen. Das bedeutet, dass der Textreader alle TIF-Dokumente, die aus ZIP-Archiven entpackt bzw. aus PDF- oder MSG-Dokumenten extrahiert werden, dort ablegt, ohne sie weiter zu bearbeiten. Der Textreader kann somit Dokumente auch anderen Anwendungen zur Verfügung stellen.

Sind zwei Werte eingetragen (siehe oben *dirs_eml*), dann extrahiert der Textreader den Text aus den Dokumenten im ersten Verzeichnis (dem sogenannten **Ausgabeverzeichnis**) und legt im zweiten Verzeichnis (dem sogenannten **Rückgabeverzeichnis**) eine TXT-Datei mit dem extrahierten Text, unter Beibehaltung des Dateinamens, ab.

Beachten Sie: Die einzelnen Konverter innerhalb des Textreaders sind teilweise zur Bearbeitung mehrerer Dateitypen programmiert worden, können aber nur einmal konfiguriert werden (insbesondere kann nur ein sogenanntes Ausgabeverzeichnis konfiguriert werden). Bei unterschiedlicher Konfiguration für einen zweiten Dateityp würde die erste Konfiguration überschrieben werden, z. B. auch das Ausgabeverzeichnis, das dann nicht mehr gescannt werden würde. Folgende Dateitypen müssen also in der *config.xml* genau gleich konfiguriert werden:

- *doc* und *dot*
- *html* und *htm*
- *jpg* und *jpeg*
- *pdf* und *ai*
- *ppt* und *pps*
- *tif* und *tiff*
- *log* und *csv*
- *xls*, *xla* und *xlt*
- *7zip* und *7z*

Sind mehr als zwei Werte eingetragen (siehe oben *dirs_pdf*), so ist ihre Bedeutung der unteren Tabelle zu entnehmen.

Eintrag in der *config.xml*	Erläuterung
`dirs_delete`	Aus diesem Verzeichnis wird alles gelöscht. Dieser Parameter kann maximal ein Zielverzeichnis enthalten.
`file_max_size_MB`	Parameter für die maximale Dokumentgröße. Die Dokumentgröße wird sowohl beim Export aus dem Archiv in die Textreader-Ordner als auch bei der Extraktion von Bilddateien aus PDF-Dateien durch die PDFBox überprüft. Dateien, deren Größe die max. Größe überschreiten, werden nach dem Export bzw. der Extraktion sofort wieder gelöscht. **Achtung**: Dieser Parameter muss unbedingt angegeben werden, da sonst keine Dateien bearbeitet werden.
`file_max_error`	Maximale Anzahl fehlerhafter Dateien, die nicht bearbeitet bzw. nicht gelöscht werden konnten. Wird die Anzahl überschritten, so stoppt der Textreader. Schützt das Überlaufen der Verzeichnisse im Fehlerfall.

`file_max_error_per_day`	Maximale Anzahl fehlerhafter Dateien pro Tag, die nicht bearbeitet bzw. nicht gelöscht werden konnten. Ist dieser Wert gesetzt, wird der Wert von *file_max_error* ignoriert. Der Zähler der nichtkonvertierten Dokumente wird am nächsten Tag wieder auf 0 gesetzt und der Textreader arbeitet wieder.
`convert not possible value`	Text, der im Fehlerfall in die TXT-Datei geschrieben wird.
`dirs_ai, dirs_bmp, dirs_csv, dirs_doc, dirs_docx, dirs_dot, dirs_dxl, dirs_eml, dirs_gif, dirs_gz, dirs_htm, dirs_html, dirs_jpg, dirs_jpeg, dirs_log, dirs_mht, dirs_mmf, dirs_msg, dirs_odt, dirs_png, dirs_ppt, dirs_pptx, dirs_rtf, dirs_vcf, dirs_vsd, dirs_vsdx, dirs_wmf, dirs_xml, dirs_xla, dirs_xls, dirs_xlsx, dirs_xlt, dirs_zip, dirs_7zip`	Diese Parameter können ein Eingangs- und ein Zielverzeichnis enthalten.
`dirs_pdf`	Parameter für PDF-Dokumente. Dieser Parameter kann bis zu 9 Werte enthalten: zunächst das Eingangs- und Zielverzeichnis, dann ein optionales *ocr*. Bei *ocr* werden die Dokumente durch eine OCR-Engine bearbeitet, die gleichzeitig den Text in den eingebetteten Grafiken extrahiert. Durch ein optional folgendes *on* werden die Dokumente, falls die OCR-Engine Fehler erzeugt, durch den Apache PDFBox-Konverter konvertiert bzw. falls der PDFBox-Konverter Fehler erzeugt, durch die OCR-Engine. Das *ocr* ist der Default-Fall. Lässt man *ocr* weg, erfolgt die Konvertierung durch den PDFBox-Konverter (nicht empfohlen bei Textreader-Versionen kleiner als 9.03.006.001, s. o.). Der Parameter *false* bedeutet, dass keine Bilder extrahiert werden. Nach dem Parameter *failover* kann mit *true* (Default) angegeben werden, dass im Fehlerfall die Konvertierung durch die OCR erfolgen soll.

	Es gibt einige weitere Parameter speziell für den PDFBox-Konverter. Nach dem Parameter *password:* kann ein Passwort eingetragen werden, falls die Dokumente verschlüsselt sind. Nach *encoding:* kann ein Character-Set angegeben werden, mit dem das Dokument erstellt worden ist (Default ist UNICODE).
`dirs_tif, dirs_tiff`	Parameter für TIF(F)-Dokumente. Dieser Parameter kann bis zu 4 Werte enthalten: das Eingangs- und das Zielverzeichnis, ein Konvertierungsprogramm und ein Timeout. Wenn kein Konvertierungsprogramm angegeben ist, erfolgt die Konvertierung über den ELO OCR-Service.
`file_image_min_size_pixel`	Minimale Breite und Höhe (in Pixel) für Grafiken, die bearbeitet werden sollen. Der Standardwert beträgt 64 Pixel. Der Check findet nur statt beim Extrahieren von Grafiken aus PDF-, EML- und MSG-Dateien.
`file_err_move_dir`	Fehlerhafte Dateien werden nicht gelöscht, sondern in das angegebene Verzeichnis verschoben, wenn es sich NICHT um Dokumente handelt, die vom Volltextdienst dem Textreader übergeben worden sind, z. B. Testdateien. Dokumente aber, die vom Volltextdienst dem Textreader übergeben worden waren, werden im Fehlerfall anders behandelt: für diese Dokumente werden im ELO Ordner *Administration¶ELO-Textreader¶Nicht konvertierte Dokumente* Referenzen angelegt (der Ordner *ELO-Textreader* hat die feste GUID "(E10E1000-E100-E100-E100-E10E10E10E24)", der Ordner *Nicht konvertierte Dokumente* die GUID "(E10E1000-E100-E100-E100-E10E10E10E25)"). Diese Ordner werden vom Textreader angelegt, wenn sie noch nicht vorhanden sein sollten.
`dirs_*`	Default-Ordner. Dokumente aus den Container-Formaten (PDF, ZIP, MSG usw.) werden im angegebenen Verzeichnis abgelegt, wenn kein Zielverzeichnis angegeben wurde.

`ocr_languages`	Über den Parameter *ocr_languages* werden die unterstützen Sprachen für die OCR-Bearbeitung festgelegt (Empfehlung: die im Serversetup verwendete Sprache, zusätzlich mindestens Englisch). Nur die tatsächlich verwendeten Sprachen eintragen, da der OCR-Service die Dokumente für sämtliche angegebenen Sprachen prüft. Eine Liste der möglichen Sprachen finden Sie weiter unten in diesem Dokument, aber auch nach der Installation unter Tomcat im Dokument *ocr_languages.html* im Tomcat-webapps-Textreader-Verzeichnis. Zusätzlich können Sie hier auch OCR-Texttypen eintragen: • *TT_Normal* (allgemeiner typographischer Text) • *TT_Typewriter* (Schreibmaschine) • *TT_Matrix* (Matrixdrucker) • *TT_OCR_A* (Spezial-OCR-Font, Monospace) • *TT_OCR_B* (Spezial-OCR-Font) • *TT_MICR_E13B* (Magnetic Ink Character Recognition, Achtung: hier auch Sprache „E13B“ auswählen) Normalerweise ist kein Eintrag eines OCR-Texttyps notwendig. **Ausnahme sind Lineprinter-Ausdrucke im Querformat**: Hier bitte bei Textreader- und ELO OCR Service-Versionen kleiner als 10 zusätzlich zu den Spracheinstellungen *TT_Matrix* angeben.
`ixurl`	Die vom OCR-Service benötigte Indexserver-URL
`username`	Name eines ELO Benutzers zum Einloggen in den Indexserver durch den OCR-Service
`passwd`	Passwort des ELO Benutzers (verschlüsselt oder nicht). Der Textreader führt keine nachträgliche Verschlüsselung durch.

`ocr_pdfPages`	Maximale Anzahl der Seiten eines Dokuments, für das der OCR-Service aufgerufen wird. Ist die Anzahl der Seiten größer als dieser Wert, und ist „on“ eingestellt als Parameter von „dirs_pdf“ (siehe oben), versucht der Textreader die Verarbeitung mit der PDFBox-Library. Empfehlung für **Textreader-Versionen bis 9.03.006.001**: möglichst hoher Wert, damit der Textreader immer den OCR-Service nutzt. Default-Wert *1000000*. Bei niedrigeren Werten, wenn also die Wahrscheinlichkeit hoch ist, dass auch die PDFBox genutzt werden wird, sollte der Textreader bis Version 9.03.006.001 immer in einem eigenen Tomcat laufen (wie oben schon erläutert). **Ab Textreader-Version 9.03.006.001**, in der die Probleme mit der PDFBox beseitigt worden waren, sind diese Vorsichtsmaßnahmen nicht mehr notwendig. PDF-Dateien sollten immer durch die PDFBox konvertiert werden. Ab **Textreader-Version 10** kann diese Option nicht mehr konfiguriert werden und entfällt, da PDF-Dateien **immer** durch die PDFBox konvertiert werden.
`ocr_timeoutseconds`	Timeout in Sekunden **innerhalb des OCR-Service** (im Gegensatz zum Parameter *ocr_threadtimeoutseconds* als max. Zeit **im Textreader**, die auf das Ergebnis der OCR gewartet wird, siehe unten). Ab OCR-Version 9.0.1.0 bezieht sich dieser Wert auf eine zu scannende Seite, in früheren Versionen auf das gesamte Dokument. Default ist 30 Sekunden. Nach Verstreichen dieser Zeit bricht die OCR die Verarbeitung der aktuellen Seite ab und in der Folge die Verarbeitung des gesamten Dokuments.

`ocr_threadtimeoutseconds`	Timeout pro Dokument in Sekunden **im Textreader.** Der Textreader wartet max. so viele Sekunden wie hier angegeben auf das Ergebnis der OCR. Nach Verstreichen dieser Zeit fährt der Textreader mit der Konvertierung des nächsten Dokuments fort (im Gegensatz zu dem Parameter *ocr_timeoutseconds* als Timeout pro Seite **innerhalb des OCR-Service**, siehe oben). Default ist 900 Sekunden (15 Minuten). Dieser Wert ist ab Version 9.02.04 konfigurierbar, da es in seltenen Fällen vorkommen kann, dass für einzelne, sehr große oder komplexe Dokumente eine Konvertierungszeit von 15 Minuten nicht ausreichend ist. Ab Version **12.00.001.000** (bzw. für TR11 ab Version 11.01.001.000, für TR10 ab Version 10.19.010.000) wird das Timeout pro Dokument berechnet **aus der Anzahl der Seiten** multipliziert mit dem Timeout pro Seite („ocr_timeoutseconds"). Nur wenn die Anzahl der Seiten nicht feststellbar ist, wird der Wert von „ocr_threadtimeoutseconds" als Default benutzt.
`ocr_files`	Anzahl der parallel zu verarbeitenden OCR-Dokumente. Hier keinen größeren Wert angeben als die in der Registry unter *WorkerCount* (z. B. unter HKEY_LOCAL_MACHINE\SOFTWARE\Wow6432Node\ELO Digital\OCR\Service) konfigurierte Anzahl von OCR-Worker-Prozessen. Default-Wert ist 2.
`ocr_fastMode`	Hier kann die Genauigkeitsrate der OCR beeinflusst werden: Beschleunigung der OCR, aber schlechtere Genauigkeitsrate durch `ocr_fastMode=true`, bzw. keine Beschleunigung, aber bessere Genauigkeitsrate durch `ocr_fastMode=false`. Default ist „false".

`ocr_singleColumnMode`	Durch *ocr_singleColumnMode* kann für die Erkennung durch die OCR festgelegt werden, dass eine Tabelle spaltenweise bearbeitet wird (`ocr_singleColumnMode = false, Default`), oder zeilenweise (`ocr_singleColumnMode = true`). Eventuell kann eine bessere Erkennung erzielt werden, wenn die OCR nicht im Single-Column-Modus läuft, also mit spaltenweiser Bearbeitung. Nachteil ist in diesem Fall aber, dass keine Abstandssuche mehr möglich ist, da die Wörter einer Zeile der Tabelle im Ergebnis nicht mehr nacheinander extrahiert werden (sondern spaltenweise).
`minimumfreediskspacepercentage`	Hier kann (in Prozent der Gesamtgröße der betreffenden Partition) angegeben werden, wieviel Plattenplatz minimal frei bleiben muss, wenn der PDFBox-Konverter Dateien extrahiert (Wert „0“ bedeutet keine Prüfung; ist kein Wert angegeben, werden Default-Werte in Abhängigkeit vom Betriebssystem benutzt).
`minimumfreediskspace`	Hier kann in MB angegeben werden, wieviel Plattenplatz minimal frei bleiben muss, wenn der PDFBox-Konverter Dateien extrahiert (Wert „0“ bedeutet keine Prüfung; ist kein Wert angegeben, werden Default-Werte in Abhängigkeit vom Betriebssystem benutzt).
`maxNumberOfFilesInFolder`	Hier kann angegeben werden, wieviel Dateien maximal in ein Verzeichnis geschrieben werden, wenn der PDFBox-Konverter Dateien extrahiert (Wert „0“ bedeutet keine Prüfung; ist kein Wert angegeben, werden Default-Werte in Abhängigkeit vom Betriebssystem benutzt).
`trace_flag`	Wird *trace_flag* auf *false* gesetzt, werden solche Konverter-Logausgaben unterdrückt, die nur angeben, dass das jeweilige Verzeichnis gescannt worden ist. Default: *false*.
`max_conv_time_pdfbox`	Die max. Zeit zur Verarbeitung einer PDF-Datei durch Apache PDFBox (exkl. Konvertierungszeit eventueller extrahierter Bilddateien). Default: 10 Minuten.

2.2 Beispiele für Einträge in der ‚config.xml'

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE properties SYSTEM http://java.sun.com/dtd/properties.dtd>
<properties>
<comment>parameters for this web application</comment>
<entry key="file_max_size_MB">20</entry>
<entry key="dirs_pdf">C:\ELOent\data\ft-elo\pdf|C:\ELOent\data\ft-
elo\txt</entry>
<entry key="dirs_mmf">C:\ELOent\data\ft-elo\mmf|C:\ELOent\data\ft-
elo\wmf</entry>
<entry key="dirs_wmf">C:\ELOent\data\ft-elo\wmf|C:\ELOent\data\ft-
elo\txt</entry>
<entry key="dirs_tif">C:\ELOent\data\ft-elo\tif</entry>
<entry key="convert not possible value">Dokument nicht bearbeitet</entry>
<entry key="file_err_move_dir">C:\ELOent\data\ft-elo\notconvfiles</entry>
<entry key="file_image_min_size_pixel">70</entry>
</properties>
```

2.2.1 Erläuterungen zur Konfiguration der Dokumentengröße

Es können nur max. 20 MB große Dokumente bearbeitet werden.

2.2.2 Erläuterungen zur Konfiguration von PDF-Dokumenten

Aus den PDF-Dokumenten wird der Text extrahiert und in einer Textdatei im Verzeichnis `C:\ELOent\data\ft-elo\txt` abgelegt. Sollte das PDF-Dokument ein TIF-Dokument enthalten, so wird dieser unter `C:\ELOent\data\ft-elo\tif` gespeichert. Alle anderen Bilddateien (JPG, PNG) werden nicht berücksichtigt, weil sie dem Textreader nicht bekannt sind. Sollen auch die übrigen Grafikformate bearbeitet werden, so ist die *config.xml* um entsprechende Entries zu ergänzen (z. B.: `<entry key="dirs_jpg">C:\ELOent\data\ft-elo\jpg</entry>`)

Der OCR-PDF-Konverter benötigt zusätzlich die Konfiguration des Dateityps *ai* mit dem Eingangsordner *pdf* (nicht etwa *ai*). Auch hier muss der Parameter *ocr* hinzugefügt werden.

Das Setzen des Parameters *on* in der Konfiguration von PDF-Dokumenten ist evtl. dann sinnvoll, wenn von vorneherein bekannt ist, dass es auch Dokumente gibt, die sich mit dem OCR-Service nicht verarbeiten lassen, weil z. B. eine Schutzeinstellung existiert, oder mit der PDFBox nicht konvertiert werden können, weil das Format anders ist als von der PDFBox erwartet (und diese *„defektes Dokument"* anzeigen würde). Durch *on* wird nach Fehlern eines der beiden PDF-Konverter ein weiterer Konvertierungsversuch durch den jeweils anderen Konverter versucht.

2.2.2.1 Kombinationen von Einstellungsmöglichkeiten bzgl. PDF-Dokumente

Es gibt also folgende Kombinationen von Einstellungsmöglichkeiten bzgl. des Konvertierungsverfahrens von PDF-Dokumenten:

a) **<entry key="dirs_pdf" C:\ELOent\data\ft-elo\pdf|C:\ELOent\data\ft-elo\txt|ocr</entry>**

Der Default-Fall. Die Verarbeitung erfolgt über OCR. PDFBox wird **nicht verwendet**, auch nicht bei Fehlern.

b) **<entry key="dirs_pdf" C:\ELOent\data\ft-elo\pdf|C:\ELOent\data\ft-elo\txt |ocr|off</entry>**

Die Verarbeitung erfolgt über OCR. PDFBox wird **nicht verwendet**, auch nicht bei Fehlern.

c) **<entry key="dirs_pdf" C:\ELOent\data\ft-elo\pdf|C:\ELOent\data\ft-elo\txt |ocr|on</entry>**

Es wird überprüft, ob es sich um ein **sehr großes Dokument** handelt (Anzahl der Seiten größer als der Wert von **ocr_pdfPages**). Wenn ja, erfolgt die Konvertierung durch den **PDFBox**-Konverter. Wenn nein, erfolgt die Verarbeitung über OCR. Der Default-Wert von *ocr_pdfPages* ist 1000000, so dass normalerweise aber immer die OCR genutzt wird (soll davon abgewichen und je nach Dokumentengröße gewechselt werden zwischen OCR-Service und PDFBox-Konverter, sollte der Textreader bei Textreader-Versionen kleiner als 9.03.006.001 wegen der Absturzgefahr des PDFBox-Konverters in einem eigenen Tomcat installiert werden). Durch das **on** wird bewirkt, dass nach **OCR-Fehlern** ersatzweise die Konvertierung durch den **PDFBox**-Konverter angestoßen wird (jetzt ohne Überprüfung der Seitenanzahl).

d) **<entry key="dirs_pdf" C:\ELOent\data\ft-elo\pdf|C:\ELOent\data\ft-elo\txt</entry>**

Ohne Zusatz von „|ocr" wird der **PDFBox**-Konverter genutzt. Bei Fehlern wird keine Konvertierung durch die OCR versucht.

e) **<entry key="dirs_pdf" C:\ELOent\data\ft-elo\pdf|C:\ELOent\data\ft-elo\txt|on</entry>**

Ohne Zusatz von „|ocr" wird der **PDFBox**-Konverter genutzt. Bei Fehlern wird die Konvertierung durch die OCR versucht.

f) **<entry key="dirs_pdf" C:\ELOent\data\ft-elo\pdf|C:\ELOent\data\ft-elo\txt|off</entry>**

Ohne Zusatz von „|ocr" wird der **PDFBox**-Konverter genutzt. Bei Fehlern wird keine Konvertierung durch die OCR versucht.

2.2.3 Erläuterungen zur Konfiguration von MMF-Dokumenten

Ein MMF-Dokument (COLD) wird in das Format WMF konvertiert und im Eingangsverzeichnis des WMF-Konverters abgelegt. Um MMF-Dokumente zu bearbeiten, muss ein WMF-Konverter unbedingt konfiguriert werden.

2.2.4 Erläuterungen zur Konfiguration von WMF-Dokumenten

Aus WMF-Dokumenten wird Text extrahiert und unter `C:\ELOent\data\ft-elo\txt` in einer Textdatei abgelegt.

2.2.5 Erläuterungen zur Konfiguration von TIF-Dokumenten

TIF-Bilddateien z. B. aus einem PDF-Dokument werden unter `C:\ELOent\data\ft-elo\tif` abgelegt, der Textreader bearbeitet diese Dateien/ Dokumente aber nicht weiter. Dies gilt für alle Formate, wenn nur ein Verzeichnis angegeben ist und somit das Zielverzeichnis fehlt.

2.2.6 Erläuterungen zur Konfiguration von 7-Zip-Dokumenten

Es werden folgende Kompressionsverfahren unterstützt:

- *AES256*
- *SHA256*
- *BZIP2*
- *BCJ_IA64_FILTER*
- *BCJ_PPC_FILTER*
- *BCJ_X86_FILTER*
- *BCJ_ARM_FILTER*
- *BCJ_ARM_THUMB_FILTER*
- *BCJ_SPARC_FILTER*
- *DEFLATE*
- *DEFLATE64*
- *DELTA_FILTER*
- *LZMA*
- *LZMA2*

7-Zip-Dokumente mit anderen Kompressionsverfahren werden als fehlerhaft behandelt.

2.2.7 Erläuterungen zur Konfiguration von minimaler Breite und Höhe

Minimale Breite und Höhe (in Pixel) für Grafiken, die bearbeitet werden sollen, beträgt hier 70 Pixel. Das ist wichtig für extrahierte Icons aus z. B. Office- oder PDF-Dateien; die Icons sollen meist nicht in den Volltext aufgenommen werden.



Beachten Sie: Der Check auf minimale Größe wird zurzeit nur bei der Extraktion von Grafiken aus PDF-, MSG- oder EML-Dokumenten durchgeführt.

2.2.8 Erläuterungen zur Konfiguration des Verhaltens im Fehlerfall

Sollte bei der Extrahierung des Textes ein Fehler passieren, wird eine Textdatei mit dem Inhalt *„Dokument nicht bearbeitet“* erzeugt.

Für fehlerhafte ELO Dateien werden durch den Textreader im ELO Ordner `Administration¶ELO-Textreader¶Nicht konvertierte Dokumente` Referenzen angelegt, siehe oben. **Nicht** ELO Dokumente werden im Fehlerfall in das Fehler-Verzeichnis geschoben, wenn ein solches in der Datei *config.xml* angegeben worden ist (*file_err_move_dir*).

3 Konfiguration über die ELO Administration Console (ab Textreader 10, beschrieben für Version 10.01.000)

Die Konfiguration des Textreaders finden Sie nach Aufruf der ELO Administration Console unter *Servermodule > Volltextdienst (Textreader)*. Die Erfahrung hat gezeigt, dass es nicht notwendig ist, in der Administration Console alle Parameter anzubieten, die in der *config.xml* konfigurierbar waren; dies führte im Gegenteil zu Konfusion bei den Administratoren. Deshalb wurden die Textreader-Konfigurationsmöglichkeiten in der Administration Console gegenüber vorigen Versionen reduziert auf sinnvoll zu ändernde Systemparameter (für die übrigen werden Default-Werte angenommen). In der Datenbank ist es aber weiter möglich, auch im Folgenden nicht aufgelistete Parameter anders zu konfigurieren, als der Textreader als Default-Wert annimmt. In der Auflistung der Default-Werte weiter unten finden Sie deshalb auch die jeweilige „optid" aus der Datenbanktabelle „eloftopt", wenn Sie Default-Werte ändern möchten.

3.1 Aufbau der ‚config.xml' ab Textreader 10

In der *config.xml* verbleiben ab Textreader 10 lediglich die URL des Indexservers sowie Login-Name und Passwort, wie man im folgenden Beispiel sieht.

Beispiel einer Datei *config.xml* ab Textreader 10:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE properties SYSTEM "http://java.sun.com/dtd/properties.dtd">
<properties>
<comment>parameters for this web application</comment>
<entry key="ixurl">http://srvtdev-elo11-1:9090/ix-elo110/ix</entry>
<entry key="passwd">xxx</entry>
<entry key="username">Administrator</entry>
</properties>
```

Die Textreader-Parameter werden in den <entry>-Tags in der *config.xml* Datei hinterlegt. Ein Parameter besteht aus einem Namen, der im Attribut *key* des <entry>-Tags eingetragen ist, und dem dazugehörigen Wert.

3.2 Bereich ‚Rückgabeverzeichnis'

Hier lässt sich das Rückgabeverzeichnis für die Textreader-Konverter einstellen (in der bisherigen *config.xml* das *Zielverzeichnis*, das nach dem ersten „|"-Symbol angegeben werden konnte für die einzelnen Dateitypen). Die Textdateien aus diesem Ordner werden als Volltextdateien in das ELO Archiv hochgeladen.

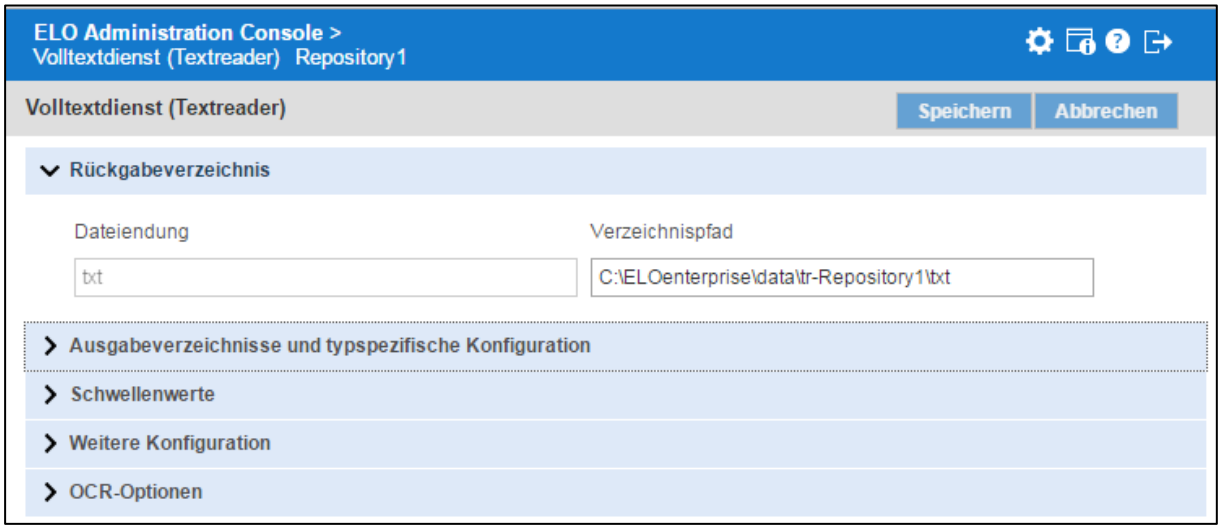


Abb. 2: Volltextdienst, Rückgabeverzeichnis

3.3 Bereich ‚Ausgabeverzeichnisse und typspezifische Konfiguration'

Unter *Ausgabeverzeichnissen* sind in diesem Fall die Verzeichnisse zu verstehen, in die die Dokumente aus dem ELO Archiv zur Konvertierung exportiert werden. Hier ist für jeden Dateityp, der in den Volltext aufgenommen werden soll, ein Verzeichnis einzustellen.

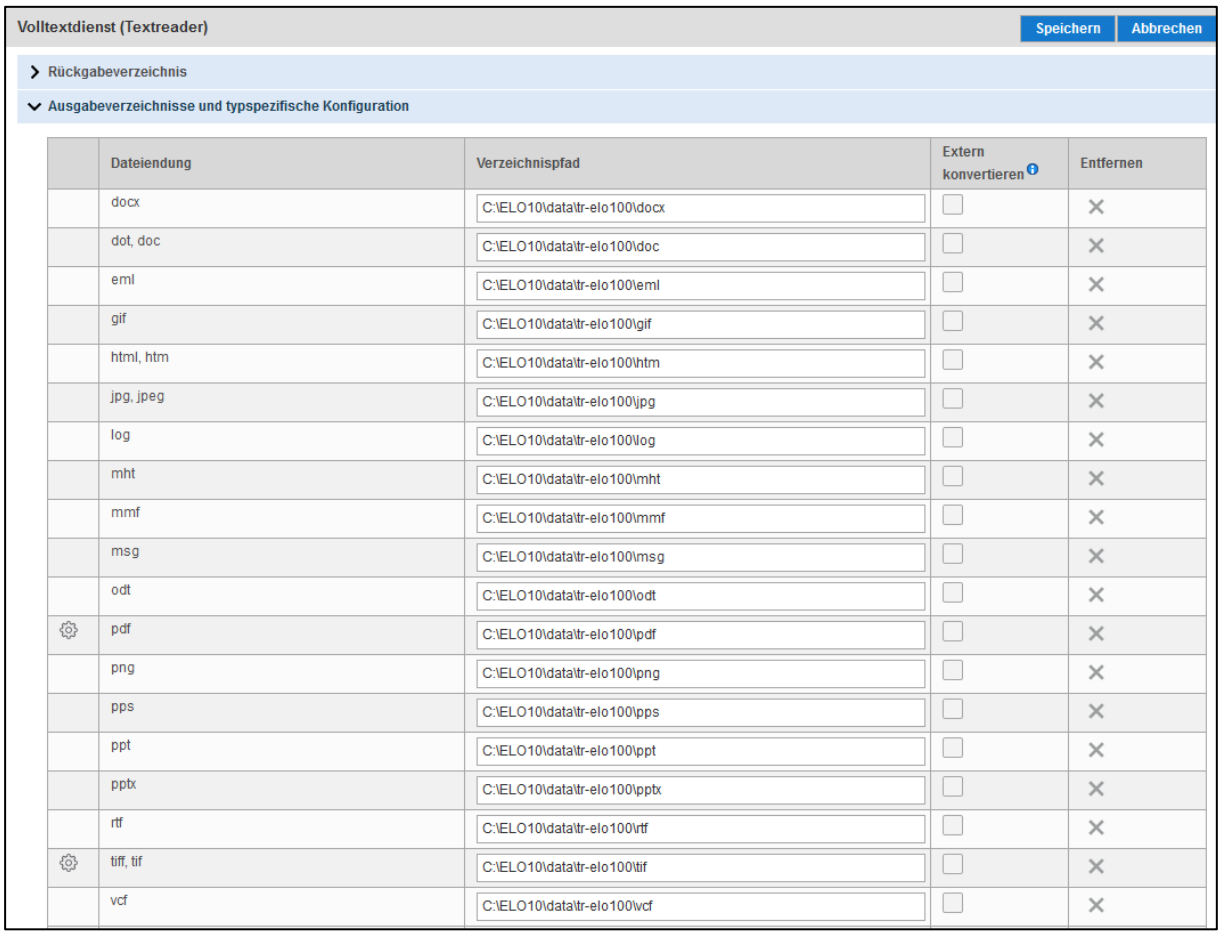


Abb. 3: Volltextdienst, Ausgabeverzeichnisse

Es ist möglich, in einer Zeile mehrere Dateitypen anzugeben, wenn diese für **einen** Verzeichnispfad identisch konfiguriert werden sollen (siehe weiter unten).

Können für einen Dateityp Optionen eingestellt werden, wird in der jeweiligen Zeile das Bearbeitungssymbol angezeigt. Nach Klick auf dieses Zahnrad-Symbol wird der Dateityp aufgeklappt und es können weitere Optionen angeklickt bzw. ausgefüllt werden. Durch einen weiteren Klick auf das Zahnrad wird die Option wieder geschlossen (aber noch nicht gespeichert, dazu müssen Sie „Speichern" anklicken!).

3.3.1 Identisch zu konfigurierende Dateitypen

Die einzelnen Konverter innerhalb des Textreaders sind teilweise zur Bearbeitung mehrerer Dateitypen programmiert worden, können aber nur einmal konfiguriert werden (insbesondere kann nur ein sogenanntes Ausgabeverzeichnis konfiguriert werden). Bei unterschiedlicher Konfiguration für einen zweiten Dateityp würde die erste Konfiguration überschrieben werden, z. B. auch das Ausgabeverzeichnis, das dann nicht mehr gescannt werden würde. Folgende Dateitypen müssen also genau gleich konfiguriert werden (bis auf die sich nur auf den Dateityp beziehende Option *Extern konvertieren*):

- *doc* und *dot*
- *html* und *htm*
- *jpg* und *jpeg*
- *pdf* und *ai*
- *ppt* und *pps*
- *tif* und *tiff*
- *xls*, *xla* und *xlt*
- *7zip* und *7z*

Es ist deshalb sinnvoll, wenn Sie diese identisch zu konfigurierenden Dateitypen jeweils in **einer** Zeile zusammenfassen, wie in unserem obigen Beispiel die Dateitypen *doc* und *dot*, *html* und *htm*, *jpg* und *jpeg*, *tiff* und *tif, 7zip und 7z*. Nur wenn Sie bestimmte dieser Dateitypen nicht durch den Textreader, sondern extern konvertieren lassen möchten, müssen Sie diese Dateitypen getrennt (d. h. in einer eigenen Zeile) konfigurieren.



Beachten Sie: Die Konfiguration folgender Dateitypen sollte jeweils in **einer** Zeile, in jedem Falle aber identisch erfolgen:

- *doc* und *dot*
- *html* und *htm*
- *jpg* und *jpeg*
- *pdf* und *ai*
- *ppt* und *pps*
- *tif* und *tiff*
- *xls*, *xla* und *xlt*
- *7zip* und 7z

3.3.2 Option ‚Extern konvertieren'

Für alle Dateitypen findet man die Option *Extern konvertieren*.

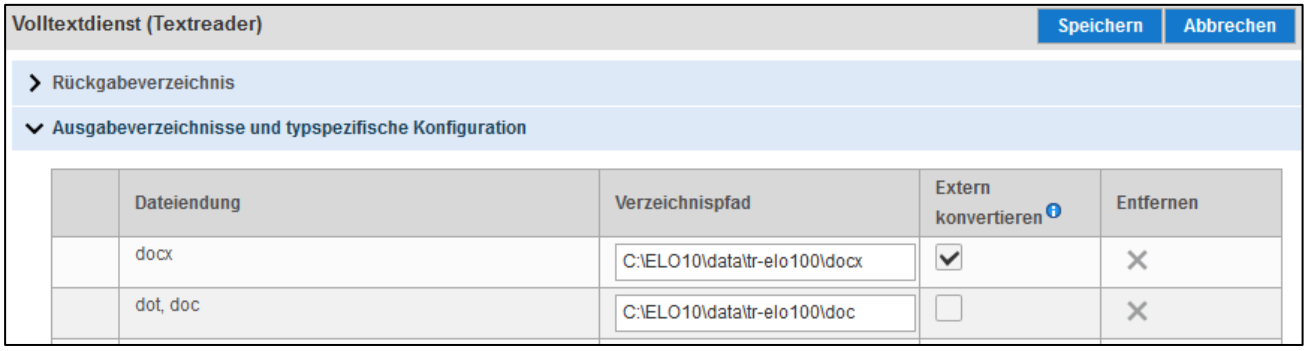


Abb. 4: Volltextdienst, Option ‚Extern konvertieren' für den Dateityp 'docx'

Erläuterung: Es müssen nicht alle Dateitypen durch die Konverter des Textreaders konvertiert werden, sondern es ist auch möglich, dass ELO Business Partner eigene Konverter einsetzen, die aber die hier beschriebenen Eingabeverzeichnisse lesen und in das Ausgabeverzeichnis schreiben. Soll also ein Dateityp durch eigene Konverter bearbeitet werden, und nicht durch die ELO Konverter, setzt man für den betreffenden Typ ein Häkchen bei *Extern konvertieren.* Das betrifft aber nur die Konverter. Der Exportservice des Textreaders würde trotzdem weiter Dokumente aus dem Archiv in solche Verzeichnisse exportieren. Der Name der Dateien im Ausgabeverzeichnis muss dem Namen der Datei im Eingabeverzeichnis entsprechen, nur insofern geändert, als dass der Dateityp nun *TXT* sein muss (statt dem ursprünglichen Dateityp). Beispiel: Die Namen der Dateien im Eingabeverzeichnis bestehen nach dem Export aus dem Archiv immer aus dem Hexadezimalwert der ELO Dokumenten-ID plus dem jeweiligen Dateityp, also z. B. „23b.docx". In diesem Fall müsste die konvertierte Datei im Ausgabeverzeichnis also heißen „23b.txt".

Auch bei mehrseitigen Dokumenten wird genau eine TXT-Datei im Ausgabeverzeichnis erwartet.

Dateien im Containerformat (also z. B. Dateityp *ZIP*, aber auch die neueren Office-Formate *DOCX*, *XLSX* usw.) werden in Einzeldateien zerlegt (Bilder und Anhänge). Die konvertierten Einzeldateien müssen jeweils einen dreistelligen Zusatz in Form eines Zählers erhalten, zum Beispiel „23b_001.txt", „23b_002.txt". Der Indexserver weiß anhand des ersten Teils des Dateinamens (der Dokumenten-ID im hexadezimalen Format), welchen ELO Dokumenten die Textdatei als Volltext zuzuordnen ist.

3.3.3 Optionen für den Dateityp ‚pdf'



Abb. 5: Volltextdienst, Optionen für den Dateityp ‚pdf'

3.3.3.1 Option ‚Bilder mit dem PDF-Konverter extrahieren'

Setzen Sie hier ein Häkchen, wenn der PDF-Konverter Bilder extrahieren soll.

3.3.3.2 Option ‚PDF-Konverter'

Wählen Sie aus der Liste den zu verwendenden PDF-Konverter (zurzeit *Apache PDFBox* oder *ICEpdf*).

3.3.3.3 Option ‚Passwort hinterlegen'

Geben Sie hier ein Passwort ein, wenn ein PDF-Dokument passwortgeschützt ist (wird vom Apache PDFBox-Konverter beim Öffnen der PDF-Datei verwendet).

3.3.4 Optionen für den Dateityp ‚tif(f)'



Abb. 6: Volltextdienst, Optionen für den Dateityp 'tif(f)'

Die Optionen in diesem Abschnitt sind nützlich, wenn kein ELO OCR Service installiert ist.

3.3.4.1 Option ‚Aus Container-Dateien nur exportieren'

Für den Fall, dass die ELO OCR nicht installiert ist, können Sie TIF(F)-Dateien aus Container-Dateien (wie z. B. ZIP-Dateien) in das TIFF-Verzeichnis extrahieren und diese durch ein separates Konvertierungsprogramm verarbeiten lassen. Aktivieren Sie hier entsprechend das Kontrollkästchen. Bei aktiviertem Kontrollkästchen durchläuft der Textreader dieses Verzeichnis nicht, kopiert aber bei der Konvertierung von z. B. ZIP-Dateien TIFF-Dateien in dieses Verzeichnis. Unterschied zu *Extern konvertieren*: Der Textreader würde, wenn die Option *Extern konvertieren* ausgewählt worden wäre, Dateien während des Export-Vorganges nicht in das TIFF-Verzeichnis verschieben.

3.3.4.2 Option ‚Externer TIFF-Konverter'

Tragen Sie hier ein externes Konvertierungsprogramm ein, das vom Textreader für die einzelnen TIFF-Dateien aufgerufen werden soll. Ansonsten erfolgt die Konvertierung über die OCR. Unterschied zu den oben beschriebenen separaten Konvertierungsprogrammen: Dieser externe TIFF-Konverter muss nicht etwa zeitgesteuert selbstständig Verzeichnisse nach zu konvertierenden Dateien durchsuchen, sondern wird unter Kontrolle des Textreaders aufgerufen und konvertiert jeweils nur ein Dokument.

3.3.4.3 Option ‚Zeitspanne für den externen TIFF-Konverter'

Hier wird ein Timeout-Wert für den externen TIFF-Konverter angegeben.

3.3.5 Konvertierung von Textdateien

Ohne dass es einer weiteren Konfiguration bedarf, werden Textdateien vom Dateityp *TXT* direkt in das Rückgabeverzeichnis exportiert und aus diesem als Volltextdateien wieder in das Archiv importiert.

Es ist aber auch möglich, Textdateien mit einem anderen, im Textreader unbekannten Dateityp in den Volltext aufzunehmen (z. B. dem Dateityp *SQL*), ohne dass dafür ein externer Konverter zur Verfügung gestellt werden muss. Diese Textdateien können ohne Konvertierung direkt als Volltextdateien verwendet werden. Dazu wird in der Konfiguration in der ELO Administration Console als Verzeichnispfad das Rückgabeverzeichnis eingetragen, normalerweise also das TXT-Verzeichnis.



Abb. 7: Volltextdienst, Konfiguration für Textdateien mit Dateityp 'vbs', 'js', oder 'sql'

3.3.6 Ein weiteres Verzeichnis hinzufügen

Um einen weiteren Dateityp hinzuzufügen, tragen Sie in der Zeile *Ein weiteres Verzeichnis hinzufügen* eine oder mehrere Dateiendungen und einen Verzeichnispfad ein und klicken auf *Hinzufügen*.

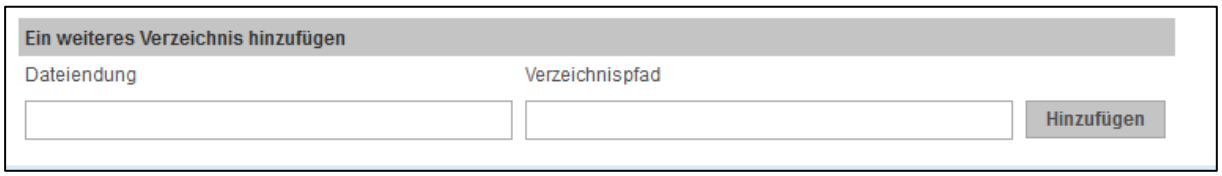


Abb. 8: Volltextdienst, Weiteres Verzeichnis hinzufügen

3.3.7 Ausgabeverzeichnisse entfernen

Sollen Dateiendungen und ihre Ausgabeverzeichnisse entfernt werden, klicken Sie auf das Entfernen-Symbol. Wird das Ausgabeverzeichnis für mehrere Dateitypen verwendet, erfolgt zunächst eine Rückfrage. Ansonsten wird die Zeile rosa eingefärbt und statt des Entfernen-Symbols ein Rückgängig-Symbol angezeigt, wie in diesem Beispiel:

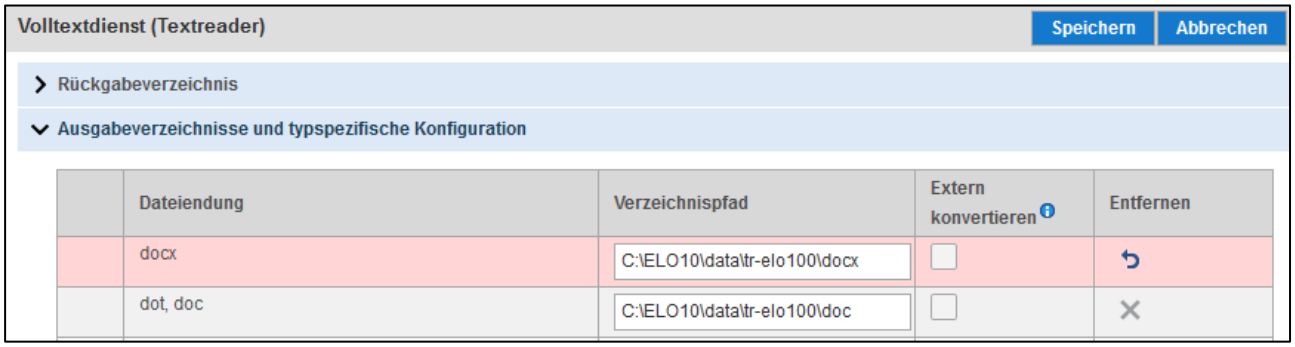


Abb. 9: Volltextdienst, Verzeichnis löschen

Ein Klick auf das Rückgängig-Symbol bewirkt, dass die Zeile wiederhergestellt wird.

3.4 Bereich ‚Textreader Schwellenwerte'

3.4.1 Option ‚Mindestbreite oder Mindesthöhe'

Minimale Breite und Höhe (in Pixel) für Grafiken, die bearbeitet werden sollen. Der Standardwert beträgt 64 Pixel.

> **Beachten Sie**: Der Check auf minimale Größe wird zurzeit nur bei der Extraktion von Grafiken aus PDF-, MSG- oder EML-Dokumenten durchgeführt.

3.4.2 Option ‚Maximale Fehlerzahl'

Maximale Anzahl fehlerhafter Dateien, die nicht bearbeitet bzw. nicht gelöscht werden konnten. Wird die Anzahl überschritten, so stoppt der Textreader. Default: 10.000.

3.4.3 Option ‚Maximal exportierte Dateien pro Minute'

Maximale Anzahl von Dateien, die pro Minute aus dem Archiv in die Eingangsverzeichnisse des Textreaders zur Konvertierung heruntergeladen werden. Bei einer größeren Anzahl von zu exportierenden Dokumenten ist es evtl. sinnvoll, hier einen Wert einzutragen, zum Beispiel 100, damit das System nicht überlastet wird und der Tomcat bzw. ELO Server zum Stehen kommt. Default: 0 (keine obere Grenze). Bei Erreichen dieser Grenze pausiert der Export-Thread im Textreader für eine Minute.



Beachten Sie: Der Textreader stellt zum Exportieren der Dokumente eine Suchanfrage an den Indexserver und erhält dann jeweils ein Paket von 100 zu exportierenden Dokumenten zurück. Muss der Textreader häufig eine Minute pausieren, weil die max. Anzahl von Dateien pro Minute erreicht ist, dauert die Bearbeitung dieser 100 Dokumente evtl. länger als die maximale Lebensdauer eines Suchtickets (Default: 10 Minuten). Wenn dann die nächsten 100 Dokumente geholt werden sollen, ist die Suche ungültig geworden und muss neu durchgeführt werden; eventuell werden Dokumente mehrfach exportiert. Sie sollten also als *maximal exportierte Dateien pro Minute* einen realistischen Wert eintragen, so dass 100 Dokumente auch in 10 Minuten exportiert werden können (bei 10 Minuten Lebensdauer sollte der Eintrag der maximal zu exportierenden Dateien pro Minute mindestens 10 betragen).

3.4.4 Option ‚Maximal importierte Dateien pro Minute'

Maximale Anzahl von konvertierten Dateien, die pro Minute wieder aus dem Rückgabeverzeichnis in das Archiv importiert werden. Bei einer größeren Anzahl von zu importierenden Dateien ist es evtl. sinnvoll, hier einen Wert einzutragen, zum Beispiel 100, damit das System nicht überlastet wird und der Tomcat bzw. ELO Server zum Stehen kommt. Default: 0 (keine obere Grenze).

3.4.5 Option ‚Maximale Fehlerzahl pro Tag'

Maximale Anzahl fehlerhafter Dateien pro Tag, die nicht bearbeitet bzw. nicht gelöscht werden konnten. Ist dieser Wert gesetzt, wird der Wert von Option ‚*Maximale Fehleranzahl*' ignoriert. Der Zähler der nichtkonvertierten Dokumente wird am nächsten Tag wieder auf 0 gesetzt und der Textreader arbeitet wieder. Default: nicht gesetzt.

3.4.6 Option ‚Maximale Dateigröße in MB'

Maximale Größe von zu konvertierenden Dateien in MB. Default 40. Die Dokumentgröße wird sowohl beim Export aus dem Archiv in die Textreader-Ordner als auch bei der Extraktion von Bilddateien aus PDF-Dateien durch die PDFBox überprüft. Dateien, deren Größe die max. Größe überschreiten, werden nach dem Export bzw. der Extraktion sofort wieder gelöscht.

3.4.7 Option ‚Zeitspanne pro Seite im OCR-Service'

Zeitspanne in Sekunden bis Timeout **innerhalb des OCR-Service** (im Gegensatz zu dem Parameter *Zeitspanne pro OCR-Dokument im Textreader* als max. Zeit **im Textreader**, die auf das Ergebnis der OCR gewartet wird, siehe unten). Ab OCR-Version 9.0.1.0 bezieht sich dieser Wert auf eine zu scannende Seite, in früheren Versionen auf das gesamte Dokument. Default ist 30 Sekunden. Nach Verstreichen dieser Zeit bricht die OCR die Verarbeitung der aktuellen Seite ab und in der Folge die Verarbeitung des gesamten Dokuments.

3.4.8 Option ‚Zeitspanne pro OCR-Dokument im Textreader'

Zeitspanne in Sekunden bis Timeout pro Dokument **im Textreader**. Der Textreader wartet max. so viele Sekunden wie hier angegeben auf das Ergebnis der OCR. Nach Verstreichen dieser Zeit fährt der Textreader mit der Konvertierung des nächsten Dokuments fort (im Gegensatz zum Parameter *Zeitspanne pro Seite im OCR-Service* als Timeout pro Seite **innerhalb des OCR-Service**). Default ist 900 Sekunden (15 Minuten). Dieser Wert ist ab Version 9.02.04 konfigurierbar, da es in seltenen Fällen vorkommen kann, dass für einzelne sehr große oder komplexe Dokumente eine Konvertierungszeit von 15 Minuten nicht ausreichend ist. Generell könnten 15 Minuten aber auch zu lang bemessen sein.



Beachten Sie: Ab Version 12.00.001.000 (bzw. für TR11 ab Version 11.01.001.000, für TR10 ab Version 10.19.010.000) berechnet sich das Timeout pro Dokument aus der Seitenanzahl des Dokuments multipliziert mit dem Timeout pro Seite. Nur wenn die Seitenanzahl nicht festgestellt werden kann, wird als Default das hier konfigurierte Timeout pro Dokument benutzt.

3.4.9 Option ‚Mindestspeicherplatz in Prozent'

Hier können Sie angeben, wieviel freier Plattenplatz (in Prozent der Gesamtgröße der Festplatte) freibleiben muss während der Export- und Konvertierungsprozesse des Textreaders (Einschränkung bzgl. Konvertierungsprozesse: nur der PDF-Konverter mit eingestellter PDF-Box überprüft in der aktuellen Version den Plattenplatz beim Extrahieren von Bildern und Attachments). Stellen Sie *Automatische Betriebssystemvorgabe* ein, wird der Mindestspeicherplatz vom Textreader festgelegt (in Abhängigkeit vom Betriebssystem). Stellen Sie *Überprüfung deaktivieren* ein, erfolgt keine Überprüfung. Sie können aber auch frei gewählte Werte eingeben (*Individuelle Eingabe)*. In diesem Fall erscheint rechts eine Eingabemöglichkeit, wo Sie in Prozenten den freizubleibenden Plattenplatz angeben können.

Nach Erreichen der eingestellten Grenze legen der Exportprozess (und bis inkl. Textreader-Version 11 auch der Konvertierungsprozess des Textreaders) eine mehrminütige Pause ein. Anschließend wird überprüft, ob wieder genügend freier Platz vorhanden ist.

Ab Textreader-Version 12 legen die Konverter keine Pause ein, sondern bei Überschreiten des Mindestspeicherplatzes wird die Konvertierung des aktuellen PDF-Dokuments abgebrochen, das gerade extrahierte Bild oder Attachment, das zur Speicherplatzüberschreitung geführt hat, wieder gelöscht, und die Konvertierung mit dem nächsten PDF fortgesetzt. Der Textreader merkt sich dieses PDF-Dokument und checkt dann in regelmäßigen Abständen, ob nun ausreichend Platz vorhanden ist. Wenn ja, wird erneut versucht, die noch nicht bearbeiteten Bilder und Attachments dieses PDFs zu extrahieren. Dadurch wird insbesondere vermieden, dass durch Extrahieren von sehr vielen großen Tiff-Dateien aus einer PDF-Datei der Textreader zum Stillstand gebracht wird.

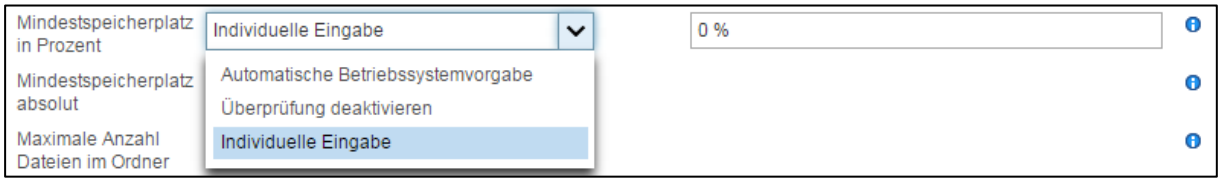


Abb. 10: Volltextdienst, Mindestspeicherplatz in Prozent

3.4.10 Option ‚Mindestspeicherplatz absolut'

Hier können Sie angeben, wieviel freier Plattenplatz (in MB) bei den Export- und Konvertierungsprozessen des Textreaders frei bleiben muss (Einschränkung bzgl. Konvertierungsprozesse: nur der PDF-Konverter mit eingestellter PDFBox überprüft in der aktuellen Version den Plattenplatz). Stellen Sie *Automatische Betriebssystemvorgabe* ein, wird der Mindestspeicherplatz vom Textreader festgelegt (in Abhängigkeit vom Betriebssystem). Stellen Sie *Überprüfung deaktivieren* ein, erfolgt keine Überprüfung. Sie können aber auch frei gewählte Werte eingeben (*Individuelle Eingabe*). In diesem Fall erscheint rechts eine Eingabemöglichkeit, wo Sie in MB den freizubleibenden Plattenplatz angeben können. Stellen Sie *Überprüfung deaktivieren* ein, erfolgt keine Überprüfung.

Nach Erreichen der eingestellten Grenze legen der Exportprozess (bzw. bis inkl. Textreader-Version 11 auch der Konvertierungsprozess des Textreaders) eine mehrminütige Pause ein. Anschließend wird überprüft, ob wieder genügend freier Platz vorhanden ist.

Ab Textreader-Version 12 legen die Konverter keine Pause ein, wie oben beschrieben.

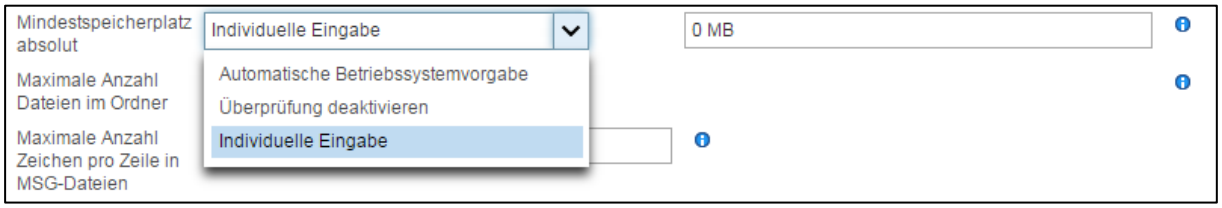


Abb. 11: Volltextdienst, Mindestspeicherplatz absolut

3.4.11 Option ‚Maximale Anzahl Dateien im Ordner' (bis Textreader-Version 11.01)

Hier können Sie die max. Anzahl von Dateien der einzelnen Ausgabe-Ordner vom System festlegen lassen, die Überprüfung deaktivieren oder die Anzahl individuell konfigurieren. Nach Erreichen dieser Grenze legt der Exportprozess (bis inkl. Textreader-Version 11 auch der PDF-Konvertierungsprozess des Textreaders) eine mehrminütige Pause ein. Anschließend wird überprüft, ob die max. Anzahl von Dateien wieder unterschritten ist.

Ab Textreader-Version 11.01 wird als Default bei Neuinstallationen die maximale Anzahl von Dateien in einem Ordner nicht mehr überprüft. Die Einstellmöglichkeit ist aus der ELO Administration Console entfernt worden. Eine Konfigurationsmöglichkeit ist aber weiterhin in der Datenbank gegeben (siehe 5.2 Abschnitt ‚Schwellenwerte').

Ab Textreader-Version 12 legen die Konverter keine Pause ein bei Überschreiten der maximalen Anzahl von Dateien in einem Ordner, sondern es wird verfahren wie bei Überschreiten des minimalen Speicherplatzes (siehe oben).

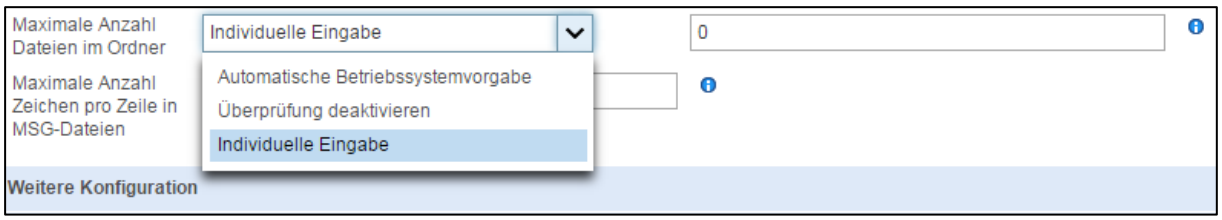


Abb. 12: Volltextdienst, Maximale Anzahl Dateien im Ordner

Beachten Sie: Ab Textreader-Version 11.01 wird bei Neuinstallationen als Default die maximale Anzahl von Dateien in einem Ordner nicht mehr überprüft.

Beachten Sie: Ab Textreader-Version 12 stoppen die Textreader-Konverter nicht mehr bei Überschreiten des minimalen Speicherplatzes oder der maximalen Datei-Anzahl in einem Ordner, sondern nur dann, wenn der Import-Prozess nicht mehr arbeitet. Die Konverter stoppen aber nicht, wenn der Import-Prozess deaktiviert worden ist.

3.4.12 Option ‚Maximale Anzahl Zeichen pro Zeile in MSG-Dateien'

Tragen Sie hier die maximale Zeilenlänge in MSG-Dateien ein. Zeilen mit einer Länge größer als dieser Wert werden bei der Konvertierung nicht berücksichtigt. Grund: es kann bei sehr langen Zeilen in MSG-Dateien zu Abstürzen des Konverters kommen. Nur aus wohlüberlegten Gründen sollte hier eine noch größere Zahl eingetragen werden. Default: 100000.

3.4.13 Option ‚Maximale Dateigröße einer Volltextdatei'

Hier können Sie die maximale Größe der resultierenden Textdatei eintragen, die als Volltextdatei wieder in das Archiv importiert wird. Default: 100 MB.

3.5 Bereich ‚Textreader weitere Konfiguration'

3.5.1 Option ‚Erweiterter Trace'

Wird die Option *Erweiterter Trace* nicht aktiviert, werden solche Konverter-Logausgaben unterdrückt, die nur angeben, dass das jeweilige Verzeichnis gescannt worden ist. Default: nicht aktiviert.

3.5.2 Option ‚Default-Ordner'

Bei der Verarbeitung von Dokumenten im Containerformat (PDF, ZIP, MSG usw.) werden aus diesen Dokumenten Dateien extrahiert und sollen je nach Dateityp in das entsprechende Zielverzeichnis kopiert werden. Ist für einzelne Dateitypen kein Zielverzeichnis konfiguriert worden, werden diese Dateien in den sogenannten Default-Ordner kopiert. Ist kein Ordner angegeben, werden solche Dateien nicht extrahiert. Default: nicht gesetzt.

3.5.3 Option ‚Freigabezeiten für Export und Import'

Hier tragen Sie (wie bisher schon beim Dienst *FT*) ein, zu welchen Tagesstunden der Export- und Importdienst zwischen dem ELO Archiv und den Textreader-Input- bzw. -Output-Verzeichnissen aktiv sein soll („00" bis „23").

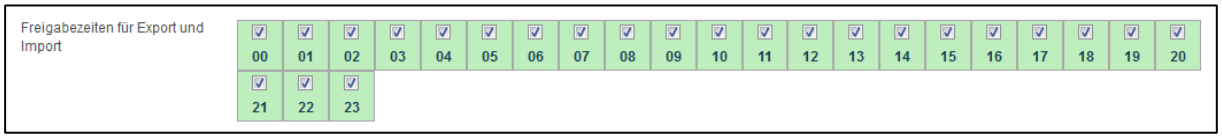


Abb. 13: Volltextdienst, Freigabezeiten für Export und Import

3.5.4 Option ‚Allgemeines Verhalten im Fehlerfall'

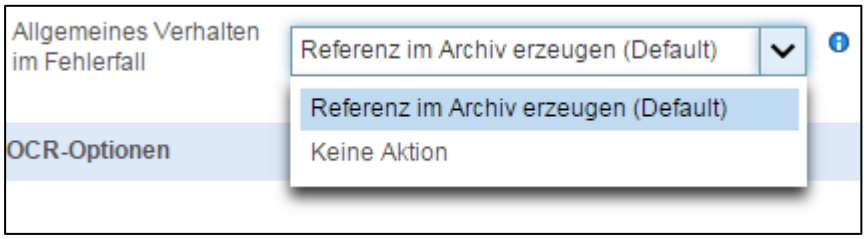


Abb. 14: Volltextdienst, Allgemeines Verhalten im Fehlerfall

Geben Sie hier an, ob bei Fehlern in der Konvertierung das ELO Dokument in einem speziellen Fehlerordner referenziert werden soll, oder ob keine Aktion stattfinden soll (dann wird das fehlerhafte Dokument nur im Ordner gelöscht). Für Dokumente, die aus dem ELO Archiv in die Eingangsordner des Textreaders exportiert worden waren, werden im Fehlerfall im ELO Ordner *Administration¶ELO-Textreader¶Nicht konvertierte Dokumente* Referenzen angelegt. Der Ordner *ELO-Textreader* hat die feste GUID "(E10E1000-E100-E100-E100-E10E10E10E24)", der Ordner *Nicht konvertierte Dokumente* die GUID "(E10E1000-E100-E100-E100-E10E10E10E25)". Diese Ordner werden vom Textreader angelegt, wenn sie noch nicht vorhanden sein sollten. Die Referenzen in den Fehlerordnern werden automatisch gelöscht, wenn in späteren Textreader-Läufen diese Dokumente doch noch konvertiert werden können. Diese ELO-Ordner werden dabei entsprechend der eingestellten Sprache benannt, zur Zeit Deutsch, Englisch oder Französisch, bzw. als Default-Sprache Englisch. Soll eine andere Default-Sprache benutzt werden, so kann man dies in der „messages_dflt.properties"-Datei einstellen (man öffnet dazu nach der Textreader-Installation im ELO Server-Verzeichnis die Textreader-jar-Datei „webapps\…\WEB-INF\lib\tr.jar" und dann die Properties-Datei „de\elo\tr\server\messages_dflt.properties").

3.5.4.1 Bemerkungen zu weiteren Konfigurationsmöglichkeiten in der Datenbank

- Fehlerhafte Dokumente, die NICHT aus ELO exportiert worden waren (z. B. Testdateien), werden, falls ein sogenannter Fehlerordner in der Datenbank konfiguriert worden ist (siehe unten), in diesen Ordner im Dateisystem verschoben. Ist kein Fehlerordner konfiguriert worden, werden solche Dateien im Fehlerfall gelöscht (Default: kein Löschordner konfiguriert).
- Im Fehlerfall wird als Default `notconv` bzw. bei Teildateien von Containerdateien (also z. B. von ZIP-Dateien) `notconv_sub` in die TXT-Datei geschrieben. Anderer Fehlertexte können in der Datenbank konfiguriert werden.
- Alternativ zu den oben unter *3.5.4 Option ‚Allgemeines Verhalten im Fehlerfall'* beschriebenen Verhalten können Sie in der Datenbank auch angeben, dass nur Fehler nur in der Ausgabedatei vermerkt werden und sonst keine Aktion stattfinden soll.
- Bei sogenannten Container-Dateien (also z. B. ZIP- oder MSG-Dateien, die andere Dateien enthalten) ist es eventuell nicht sinnvoll, im Fehlerfall die ganze Container-Datei im Archiv als fehlerhaft zu referenzieren. Deshalb kann hier für Container-Dateien ein vom oben ausgewählten Verhalten abweichendes Verhalten in der Datenbank konfiguriert werden.

  (Die Konfigurationsoptionen in der Datenbank sind weiter unten beschrieben.)

3.6 Bereich ‚OCR-Optionen'

3.6.1 Option ‚OCR-Lesemodus'

Durch die Option *OCR-Lesemodus* kann für die Erkennung durch die OCR festgelegt werden, dass eine Tabelle spaltenweise bearbeitet wird (Default: *Mehrspaltenmodus*, die OCR versucht Blöcke zu erkennen), oder zeilenweise (*Einspaltenmodus*, die OCR versucht nicht mehr, Blöcke zu erkennen). Eventuell kann eine bessere Erkennung erzielt werden, wenn die OCR nicht im Einspaltenmodus läuft, sondern mit spaltenweiser Bearbeitung im Mehrspaltenmodus. Nachteil ist in diesem Fall aber, dass keine Abstandssuche mehr möglich ist, da die Wörter einer Zeile der Tabelle im Ergebnis nicht mehr nacheinander extrahiert werden (sondern spaltenweise).

3.6.2 Option ‚OCR-Erkennungsmodus ausführlich'

Hier kann die Genauigkeitsrate der OCR beeinflusst werden: Beschleunigung der OCR, aber schlechtere Genauigkeitsrate durch Auswahl von *Schnell,* bzw. keine Beschleunigung, aber bessere Genauigkeitsrate durch Auswahl von *Ausführlich*. Default ist *Ausführlich*.

3.6.3 Option ‚OCR-Sprachen'

Über den Parameter *OCR-Sprachen* werden die unterstützten Sprachen für die OCR-Bearbeitung festgelegt (Empfehlung: die im Serversetup verwendete Sprache, zusätzlich mindestens Englisch). Nur die tatsächlich verwendeten Sprachen eintragen, da der OCR-Service die Dokumente für sämtliche angegebenen Sprachen prüft. Auf der linken Seite finden Sie die Liste der von der installierten OCR unterstützten Sprachen. Zusätzlich können Sie hier OCR-Texttypen auswählen, was normalerweise aber nicht erforderlich ist (insbesondere ist es ab Textreader 10 auch nicht mehr notwendig, zusätzlich zu den Spracheinstellungen *TT_Matrix* anzugeben, wenn Lineprinter-Ausdrucke im Querformat konvertiert werden sollen). Der Vollständigkeit halber seien die OCR-Texttypen hier noch einmal aufgelistet:

- *TT_Normal* (allgemeiner typographischer Text)
- *TT_Typewriter* (Schreibmaschine)
- *TT_Matrix* (Matrixdrucker)
- *TT_OCR_A* (Spezial-OCR-Font, Monospace)
- *TT_OCR_B* (Spezial-OCR-Font)
- *TT_MICR_E13B* (Magnetic Ink Character Recognition, Achtung: hier auch Sprache „E13B" auswählen)

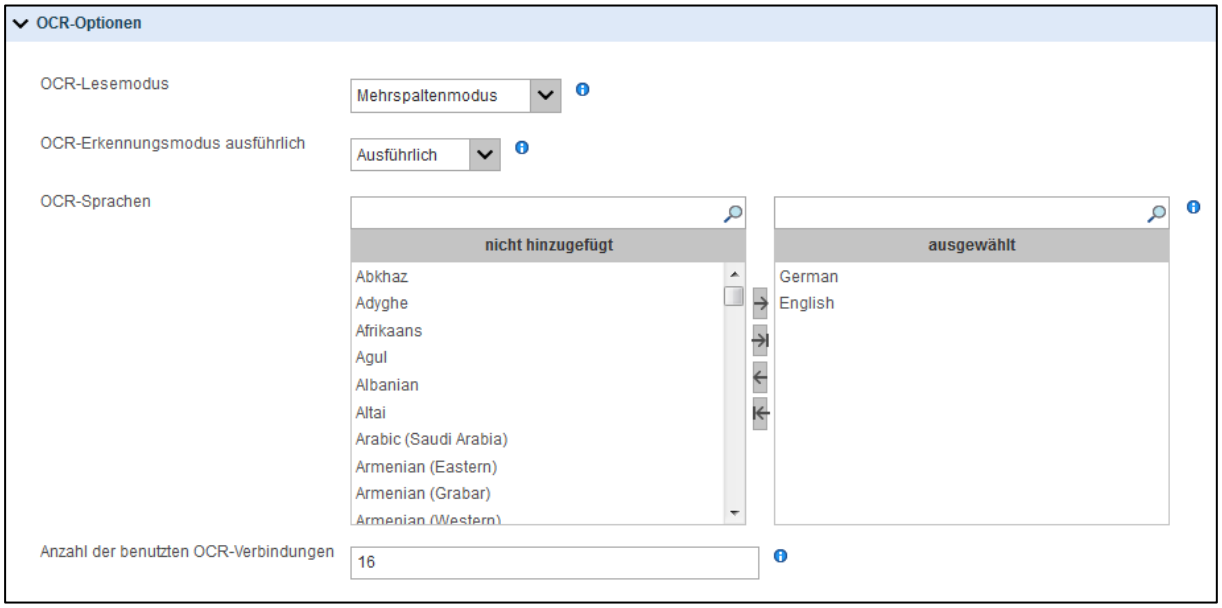


Abb. 15: Volltextdienst, Konfiguration der OCR-Sprachen

Information: Bis Version 10.00.010 der ELO Administration Console gibt es keine Möglichkeit, OCR-Texttypen zu konfigurieren. Sollen trotzdem OCR-Texttypen konfiguriert werden, tragen Sie diese bitte in der Datenbanktabelle *eloftopt* in der Zeile mit der *optid* 1139 ein.

3.6.4 Option ‚Anzahl der benutzten OCR-Verbindungen'

Hintergrund: In der OCR wird eine Anzahl von parallel arbeitenden Worker-Prozessen konfiguriert, die die eigentliche Konvertierung von Bildern durchführen. Diese Anzahl wird für OCR 9 in der Registry unter „WorkerCount" definiert, z. B. unter `HKEY_LOCAL_MACHINE\SOFTWARE\Wow6432Node\ELO Digital\OCR\Service`, bzw. für OCR 10 in der *config.xml* der OCR durch den Key „workercount". Normalerweise wird als Anzahl der OCR-Worker die Anzahl der Kerne * 2 angegeben. Die OCR erzeugt dann eine entsprechende Anzahl von Worker-Prozessen. Sie können in der Administration Console konfigurieren, wie viele von diesen Workern maximal vom Textreader benutzt werden sollen (die restlichen werden von den ELO Java Clients benutzt). Der Textreader erstellt dann entsprechend viele Threads, die wiederum (parallel) Bild-Dateien zur Konvertierung an die Worker schicken.

Geben Sie mehr Worker an, als die OCR zur Verfügung stellt, wird die OCR an den Textreader evtl. zurückmelden, dass gerade keine freien Worker zur Verfügung stünden. Der Textreader wartet in solchen Fällen etwas und versucht dann erneut, das fragliche Dokument an die OCR zu schicken. Er versucht das zwar so lange, bis ein freier Worker gefunden worden ist, bzgl. der Performanz würde es aber keine Vorteile bringen, an dieser Stelle mehr Worker zu konfigurieren als überhaupt zur Verfügung stehen.

Tragen Sie also unter *Anzahl der benutzten OCR-Verbindungen* die maximale Worker-Anzahl ein, die vom Textreader benutzt werden können, geben aber keinen größeren Wert an als in der Registry unter „WorkerCount“ bzw. in der *config.xml* der OCR unter „workercount“ konfiguriert, und ziehen evtl. eine bestimmte Anzahl von für ELO Java Clients bestimmte Worker ab.

Default-Wert ist 2.



Information: Die übrigen bisherigen FT-Optionen können entfallen, da der integrierte Textreader nun über den Indexserver die Datenbank aufruft bzw. Volltextdateien in das ELO Archiv hochlädt und nur noch ELO iSearch unterstützt wird.

4 Default-Werte der Konfigurationsoptionen nach Neuinstallation der aktuellen Textreader-Version (ab Version 10)

Nach einer Neuinstallation werden, wie oben schon erwähnt, die Ein- und Ausgabeverzeichnisse des Textreaders konfiguriert. Ansonsten arbeitet der Textreader mit den hier aufgeführten Default-Werten, die evtl. nach der Installation über die Administration Console angepasst werden sollten an Ihre speziellen Anforderungen (insbesondere die Anzahl der OCR-Worker ist anzupassen an die für den ELO OCR Service konfigurierte Worker-Anzahl).



Information: Die ELO Administration Console ergänzt fehlende Konfigurationseinstellungen durch Default-Werte (in der Datenbanktabelle „eloftopt"). Dieses Feature kann auch genutzt werden, um leere „eloftopt"-Tabellen zu füllen, z. B. nach einem Setup oder Update, bei dem die Tabelle zwar angelegt wurde, aber fälschlicherweise leer geblieben ist.

4.1.1 Abschnitt ‚Ausgabeverzeichnisse und typspezifische Konfiguration'

- Folgende Dateitypen werden konfiguriert:

  ai, *bmp*, *doc*, *docx*, *dot*, *eml*, *htm*, *html*, *jpg*, *jpeg*, *mht*, *mmf*, *msg*, *odt*, *pdf*, *png*, *pps*, *ppt*, *pptx*, *rtf*, *tif*, *tiff*, *vcf*, *vsd*, *vsdx*, *wmf*, *xla*, *xls*, *xlsx*, *xml*, *zip*, *7zip*, *7z*

- Alle Dateitypen werden bearbeitet.
- Dateityp *pdf*
  - Option *Bilder mit dem PDF-Konverter extrahieren* ist ausgewählt
  - Option *PDF-Konverter: Apache PDFBox*
  - Option *Zeichensatz: UNICODE* (nur in der Datenbank änderbar)

4.1.2 Abschnitt ‚Textreader Schwellenwerte'

- Option *Mindestbreite oder Mindesthöhe*: 64 (Pixel)
- Option *Maximale Fehlerzahl*: 10.000
- Option *Maximal exportierte Dateien pro Minute*: 0 (unbegrenzt)
- Option *Maximal importierte Dateien pro Minute*: 0 (unbegrenzt)
- Option *Maximale Fehlerzahl pro Tag*: nicht gesetzt
- Option *Maximale Dateigröße in MB*: 40
- Option *Zeitspanne pro Seite im OCR-Service*: 30
- Option *Zeitspanne pro OCR-Dokument im Textreader*: 900 Sekunden (15 Minuten)

- Optionen *Mindestspeicherplatz in Prozent, Mindestspeicherplatz absolut:* die Werte werden systemabhängig ermittelt
- *Maximale Anzahl Dateien im Ordner*: deaktiviert ab Textreader 11.01
- Option *Maximale Anzahl Zeichen pro Zeile in MSG-Dateien*: 100000

4.1.3 Abschnitt ‚Textreader weitere Konfiguration'

- Option *Zeichensatz*: UNICODE (nur in der Datenbank änderbar)
- Option *Erweiterter Trace:* nicht ausgewählt
- Option *Verschieben durch Kopieren und Löschen*: nicht gesetzt
- Option *Default-Ordner*: nicht gesetzt
- Option *Freigabezeiten für Export und Import*: stündlich („00" – „23")
- Option *Allgemeines Verhalten im Fehlerfall*: Referenz im Archiv erzeugen

4.1.4 Abschnitt ‚OCR-Optionen'

- Option *OCR-Lesemodus*: Mehrspaltenmodus
- Option *OCR-Erkennungsmodus*: Ausführlich
- Option *OCR-Sprachen*: German,English,French
- Option *Anzahl der benutzten OCR-Verbindungen*: 2

4.1.5 Abschnitt ‚Verhalten bei Fehlern'

- Option *Fehlerkennzeichnung in der Ausgabedatei*: notconv|notconvsub
- Option *Allgemeines Verhalten im Fehlerfall*: Referenz im Archiv erzeugen
- Option *Abweichungen für Container-Dateien*: Referenz im Archiv erzeugen

5 Anpassung von Default-Werten in der Datenbank (ab Version 10)

Wie schon beschrieben, werden ab Version 10 teilweise Default-Werte gesetzt, von denen angenommen werden kann, dass diese nicht verändert werden müssen. Deshalb besteht für diese Werte keine Konfigurationsmöglichkeit mehr in der ELO Administration Console. Trotzdem soll anschließend erläutert werden, wie auch diese Werte in der Datenbank-Tabelle *eloftopt* geändert werden können.



Beachten Sie: Änderungen in der Datenbank sind immer kritisch und können zu einem Stillstand des Systems führen. Ändern Sie die Werte in der Datenbank nur nach reiflicher Überlegung und wenn Sie sich über die Folgen im Klaren sind.

5.1 Abschnitt ‚Ausgabeverzeichnisse und typspezifische Konfiguration'

- Dateityp *pdf*:
  - Option *Zeichensatz*: Default ist *UNICODE.* In der Datenbanktabelle *eloftopt* änderbar bspw. unter optid *131415.* Dabei ist die *1314* die (veränderliche) optid für den Dateityp *pdf*, die Kennzeichnung für die Option *Zeichensatz* ist die *15.* (Die optid-Werte werden für die Dateitypen fortlaufend vergeben und sind deshalb variabel; die optid-Werte der Optionen einzelner Dateitypen berechnen sich aus der optid des Dateityps mit einer zur jeweiligen Option gehörenden Kennzeichnung als Anhang, hier „15".)
  - Option *Max. PDF-Konvertierungszeit*: Der Eintrag erfolgt in Minuten, Default ist 10 Minuten. In der Datenbanktabelle *eloftopt* bspw. änderbar unter optid *131417,* wenn *1314* die (veränderliche) optid für den Dateityp *pdf* ist. Die Kennzeichnung für die Option *Max. PDF-Konvertierungszeit* ist die *17.*
  - Option *Maximale Anzahl von extrahierten Graphiken pro Dokument*: Als Default werden pro Dokument maximal 200.000 Graphiken extrahiert. Anschließend läuft die Konvertierung des PDF-Dokumentes zwar weiter, aber Bilder werden nicht mehr extrahiert. In der Datenbanktabelle *eloftopt* bspw. änderbar unter optid *131419,* wenn *1314* die (veränderliche) optid für den Dateityp *pdf* ist. Die Kennzeichnung für die Option *Max. Anzahl von extrahierten Graphiken pro Dokument* ist die *19.*

- Option *Konvertierung durch die OCR im Fehlerfall („Failover“)*: Als Default erfolgt die Konvertierung von PDF-Dokumenten nur durch die PDFBox (oder, wenn so konfiguriert, durch ICEpdf). Ab Textreader-Version 12.00.005.000 (bzw. für TR11 ab Version 11.01.006.000, für TR10 ab Version 10.19.100.000) werden nach Konvertierungsfehlern die Dokumente an die OCR weitergeleitet. Dieses Feature kann deaktiviert werden durch den Eintrag von „false“ unter der optid *131418* in der Datenbanktabelle *eloftopt*, wenn *1314* die (veränderliche) optid für den Dateityp *pdf* ist. Default ist „true“ – das Feature ist also aktiviert.

  > **Beachten Sie**: Nach Konvertierungsfehlern durch die PDFBox oder ICEpdf werden PDF-Dateien als Default an die OCR geschickt.

5.2 Abschnitt ‚Schwellenwerte‘

- Option *Maximale Anzahl Dateien im Ordner*: Hier können Sie die max. Anzahl von Dateien der einzelnen Ausgabe-Ordner vom System festlegen lassen, die Überprüfung deaktivieren oder die Anzahl individuell konfigurieren. Nach Erreichen dieser Grenze legt der Export- bzw. der Konverter-Prozess des Textreaders eine mehrminütige Pause ein. Anschließend wird überprüft, ob die max. Anzahl von Dateien wieder unterschritten ist (Einschränkung bzgl. Konvertierungsprozesse: nur der PDF-Konverter mit eingestellter PDFBox überprüft in der aktuellen Version die Anzahl der Dateien). Als Default ist die Einstellung deaktiviert. Zur Aktivierung gehen Sie zur optid *1062* und tragen die max. Anzahl Dateien im Ordner ein. Die Deaktivierung würde durch Eintrag von „0“ erfolgen, die automatische Einstellung von Betriebssystemvorgaben durch Eintrag von „-1“.

  Ab Textreader-Version 12 legen die Konverter keine Pause ein (wie oben beschrieben unter *3.4.9 Option ‚Mindestspeicherplatz in Prozent‘*).

5.3 Abschnitt ‚Textreader weitere Konfiguration‘

- Option *Zeichensatz:* Default ist *UNICODE*. In der Datenbanktabelle *eloftopt* änderbar unter optid *1145*.

  > **Beachten Sie**: Für den Dateityp *pdf* geben Sie den Zeichensatz unter optid *131415* an (wenn *1314* die optid für den Dateityp *pdf* ist).

- Option *Verschieben durch Kopieren und Löschen*: Zum Aktivieren dieser Option tragen Sie in der Datenbanktabelle *eloftopt* unter optid *1146 true* ein. Default: *false*.
- Option *Löschordner:* Als Default ist dieser Ordner nicht gesetzt. In der Datenbanktabelle *eloftopt* änderbar unter optid *1148*.

- Option *Allgemeines Verhalten im Fehlerfall:* Zusätzlich zu den in der Administration Console angebotenen Alternativen *Referenz im Archiv erzeugen* und *Keine Aktion* können in der Datenbanktabelle *eloftopt* unter der optid *1149* weitere mögliche Aktionen konfiguriert werden. Hier eine vollständige Auflistung der möglichen Werte:
  - *Referenz in Archiv erzeugen*: Setzen Sie optval auf *1*.
  - *In Fehlerordner verschieben*: Setzen Sie optval auf *2*.
  - *Fehler in Ausgabedatei vermerken*: Setzen Sie optval auf *4*.
  - *Keine Aktion*: Setzen Sie optval auf *8*.
- Option *Abweichungen für Container-Dateien:* Als Default ist keine von *Allgemeines Verhalten im Fehlerfall* abweichende Aktion gesetzt. Änderbar ist dies auch in der Datenbanktabelle *eloftopt* unter optid *1149*. Um sowohl die Optionen *Allgemeines Verhalten im Fehlerfall und Abweichungen für Container-Dateien* unter einer optid zusammenfassen zu können, werden die durchzuführenden Aktionen durch Setzen von Bits in einem Bitmuster konfiguriert; der Dezimalwert dieses Bitmusters wird in der Datenbank unter optid *1149* gespeichert. Im Folgenden finden Sie eine Auflistung der möglichen Aktionen und den daraus resultierenden Werten für optid *1149*:
  - Option *Referenz in Archiv erzeugen* gesetzt für *Allgemeines Verhalten im Fehlerfall:*
    - *In Fehlerordner verschieben* soll abweichend davon als Option für Container-Dateien gesetzt werden: Setzen Sie optval auf *513.*
    - *Keine Aktion* soll als Option für Container-Dateien gesetzt werden: Setzen Sie optval auf *1025*.
  - *In Fehlerordner verschieben* gesetzt für *Allgemeines Verhalten im Fehlerfall* und *Keine Aktion* soll als Option für Container-Dateien gesetzt werden: Setzen Sie optval auf *1026*.
  - *Fehler in Ausgabedatei vermerken* gesetzt für *Allgemeines Verhalten im Fehlerfall* und *Keine Aktion* soll als Option für Container-Dateien gesetzt werden: Setzen Sie optval auf *1028*.
- Option *Fehlerordner:* Als Default ist dieser Ordner nicht gesetzt. In der Datenbanktabelle *eloftop*t änderbar unter optid *1144*.
- Option *Fehlerkennzeichnung in der Ausgabedatei*: Als Default ist *notconv|notconv_sub* eingetragen. In der Datenbanktabelle *eloftopt* änderbar unter optid *1138*.

5.4 Übersicht für Optionen in der Datenbank

Option	optid	Wert
Fehlerkennzeichnung in der Ausgabedatei	1138	Default: notconv|notconv_sub
Fehlerordner	1144	Default: nicht gesetzt
Zeichensatz für sonst. Konverter	1145	Default: UNICODE
Verschieben durch Kopieren und Löschen	1146	true/false, Default: false
Löschordner	1148	Default: nicht gesetzt
Allg. Verhalten im Fehlerfall	1149	1 (Referenz im Archiv erzeugen, auch für Containerdateien) 513 (Referenz im Archiv erzeugen, aber davon abweichend für Containerdateien: diese in Fehlerordner verschieben) 1025 (Referenz im Archiv erzeugen, aber davon abweichend für Containerdateien: keine Aktion) 2 (In Fehlerordner verschieben, auch für Containerdateien) 1026 (In Fehlerordner verschieben, aber davon abweichend für Containerdateien: keine Aktion) 4 (Fehler in Ausgabedatei vermerken, auch für Containerdateien) 1028 (Fehler in Ausgabedatei vermerken, aber davon abweichend für Containerdateien: keine Aktion) 8 (Keine Aktion, auch für Containerdateien)
Zeichensatz für PDFBox-Konverter	z. B. 131415, wenn 1314 die optid für pdf ist	Default: UNICODE
Max. Konvertierungszeit für PDFBox-Konverter	z.B. 131417, wenn 1314 die optid für pdf ist	Default: 10 Minuten