



Diese Anleitung ist ein Entwurf und noch nicht fertig! Bitte verwenden Sie diese Anleitung noch nicht!

# docsearch-Plugin: Dokumente durchsuchbar machen

Mit dem docsearch-Plugin können Dokumente, die als Medien-Dateien im Wiki hinterlegt sind, durch die Volltextsuche durchsuchbar gemacht werden (z.B. Word, Excel, Powerpoint und PDF-Dateien).

 [Dokumentation des docsearch-Plugins auf dokuwiki.org](#)



Im Gegensatz zu den meisten anderen DokuWiki-Plugins ist es beim docsearch-Plugin nicht ausreichend, einfach nur das Plugin zu installieren. Es sind noch weitergehende Konfigurationen auf der Betriebssystemebene sowie die Installation von Drittprogrammen notwendig!

Dies bedingt ein Administrator-Konto. Wird das Wiki auf einem normalen Webhosting-Paket betrieben, wird die Installation mit großer Wahrscheinlichkeit nicht durchführbar sein.

## Funktionsweise des docsearch-Plugins

Die DokuWiki-Software kann standardmäßig nur den [Suchindex](#) für Wiki-Seiten erzeugen, nicht aber für Dateien.

Das Plugin erreicht technisch gesehen auf folgendem Weg, dass auch Mediendateien durchsucht werden können:

1. Starten des Programms durch eine zeitgesteuerte, automatisierte Betriebssystemfunktion (*cronjob* in Linux, *geplante Aufgabe* in Windows)
2. Das Programm erstellt nun eine Liste von allen Mediendateien im Wiki
3. Für die Dateitypen, für die ein *converter* definiert ist (anhand der Dateinamensendung), wird ein Drittprogramm (z.B. *LibreOffice*) aufgerufen, welches die Mediendatei in reinen Text umwandelt
4. Die so erzeugte Textdatei kann nun von der DokuWiki-eigenen Funktion zur Erstellung des Suchindexes verarbeitet werden
5. Die Inhalte der Mediendateien werden in den Suchergebnissen angezeigt

## Grenzen des docsearch-Plugins

Da die Erzeugung des Suchindexes für Mediendateien ebenfalls auf dem Auslesen von maschinenlesbaren Inhalten basiert, können nur solche Dateien in den Suchindex aufgenommen werden, die auch lesbaren Text enthalten. Ein PDF, welches ein eingescanntes Dokument, also ein Bild enthält, kann durch die Suche nicht indiziert werden. Hier wäre zusätzlich die Nutzung einer Texterkennungssoftware notwendig. Dies verkompliziert die Konfiguration jedoch weiter und sorgt für einen vielfachen Ressourcenverbrauch.

## Installation des Plugins

Dieses Plugin ist in der Standard-Einsatzleiterwiki-Installation nicht enthalten. Installieren Sie das Plugin, wie in der Anleitung [Plugins installieren](#) beschrieben. In den Suchergebnissen des Plugin-Managers wird das Plugin als *Document Search Plugin* bezeichnet.

## Konfiguration des Plugins und Installation von Drittprogrammen in Linux-Betriebssystemen



Die hier beschriebenen Schritte gelten für Ubuntu-/Debian-basierte Linux-Installationen.



Bei anderen Distributionen können für die Installation andere Befehle oder Paketnamen notwendig sein!

Zum Auslesen von PDF-Dateien wird das Paket `poppler-utils` benötigt. Für Dokumente der Office-Formate, also sowohl Microsoft Office als auch OpenOffice/LibreOffice, kann das LibreOffice-Paket verwendet werden. Installieren Sie die benötigten Pakete mit folgendem Befehl (für Debian/Ubuntu und verwandte Distributionen):

```
sudo apt-get install poppler-utils libreoffice
```

Legen Sie nach der Installation der Hilfsprogramme (ausgehend von Ihrem Wiki-Stammverzeichnis) im Verzeichnis `lib/plugins/docsearch/conf` die Datei `converter.sh` an (diese existiert noch nicht). Kopieren Sie nun den folgenden Inhalt in die Datei, alternativ können Sie diese hier auch durch Klick auf den Dateinamen herunterladen und direkt in das genannte Verzeichnis speichern:

[converter.sh](#)

```
# übergebene Argumente in Variablen schreiben
INPUTFILE="$1"
OUTPUTFILE="$2"

#Pfad (ohne Dateiname) für die Ausgabedatei ermitteln
OUTPUTPATH=`dirname "$OUTPUTFILE"`

# Dateiendung der Eingabedatei ermitteln
INPUTFILEEXT=${INPUTFILE##*.}

# je nach Dateityp den entsprechenden Befehl ausführen
# siehe dazu auch
https://de.wikipedia.org/wiki/Liste_der_Microsoft-Office-Dateinamenserweiterungen
case $INPUTFILEEXT in
    pdf)
        pdftotext -enc UTF-8 $INPUTFILE $OUTPUTFILE
        ;;
    doc|dot|docx|docm|dotx|dotm|xls|xlm|xlt|xlsx|xlsb|xltm|xltm|ppt|pot|pps|pptx|pptm|potx|potm|ppsx|ppsm|odt|ott|ods|ots|csv|odp|otp|odg)
        libreoffice --headless --convert-to "txt:Text (encoded):UTF8" --outdir "$OUTPUTPATH" "$INPUTFILE"
        ;;
esac
```

Definieren Sie nun, dass die Datei bzw. das Script als Programm ausgeführt werden darf. Da dafür vermutlich root-Rechte notwendig sind, verwenden Sie folgenden Befehl. Achten Sie darauf, den Pfad entsprechend an Ihre Ordnerstruktur anzupassen, der folgende Pfad wird nur beispielhaft verwendet:

```
sudo chmod 755
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
```

Legen Sie ebenfalls im Verzeichnis `lib/plugins/docsearch/conf` eine weitere Datei

converter.php an (diese existiert auch noch nicht). Kopieren Sie nun den folgenden Inhalt in die Datei, alternativ können Sie diese hier auch durch Klick auf den Dateinamen herunterladen und direkt in das genannte Verzeichnis speichern:

### converter.php

```
#<?php die(); ?>
pdf
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
doc
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
dot
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
docx
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
docm
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
dotx
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
dotm
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
xls
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
xlm
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
xlt
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
xlsx
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
xlsm
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
xltx
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
xltn
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
ppt
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
```

```
%in% %out%
pot
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
pps
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
pptx
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
pptm
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
potx
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
potm
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
ppsx
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
ppsm
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
odt
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
ott
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
ods
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
ots
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
csv
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
odp
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
otp
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
odg
/var/www/html/einsatzleiterwiki/lib/plugins/docsearch/conf/converter.sh
%in% %out%
```

Passen Sie auch hier den Pfad in jeder Zeile an Ihre Ordnerstruktur an.

Möchten Sie bestimmte Dateitypen von der Aufnahme in den Suchindex ausschließen, so löschen Sie die entsprechende Zeile, welche mit der Dateiendung beginnt, aus der Datei `converter.php` heraus.