

Innerhalb von Dateien suchen

Sollen die Inhalte von Dateien aus den Dateiarchiven von der Suchfunktion durchsucht werden können, und hast Du ein Skript, welches den jeweiligen Dateiinhalt in einen Text extrahiert, dann kannst Du dieses Skript mit dem MIME-Typ verknüpfen und die Dateiinhalte werden sodann indexiert.

Sollen Dateien in den Dateiarchiven durchsucht werden, musst Du den Handler zum Extrahieren des Textes aus den Dateien angeben. Die Befehle, wie z.B. *strings* oder *pdftotext* müssen dabei auf Deinem Server installiert sein. Der jeweiligen Typ-Befehl muss im Indexierungs-Tab unter Dateiarchive verwalten definiert werden.

MIME Typ	Systembefehl	Ubuntu/Debian Paket mit Befehl
application/vnd.oasis.opendocument.presentation	odt2txt %1	odt2txt
application/vnd.oasis.opendocument.spreadsheet	odt2txt %1	odt2txt
application/vnd.oasis.opendocument.text	odt2txt %1	odt2txt
application/ms-excel	xls2csv %1	catdoc
application/ms-powerpoint	catppt %1	catdoc
application/msword	catdoc %1 oder strings %1	catdoc
application/pdf	pstotext %1 oder pdftotext %1 -	pstotext
application/postscript	pstotext %1	pstotext
application/ps	pstotext %1	pstotext
application/rtf	catdoc %1	catdoc
application/sgml	col -b %1 oder strings %1	bsdmainutils
application/vnd.ms-excel	xls2csv %1	catdoc
application/vnd.ms-powerpoint	catppt %1	catdoc
application/x-msexcel	xls2csv %1	catdoc

application/x-pdf	pstotext %1	pstotext
application/x-troff-man	man -l %1	man-db
text/enriched	col -b %1 oder strings %1	bsdmainutils
text/html	elinks -dump -no-home %1	elinks
text/plain	col -b %1 oder strings %1	bsdmainutils
text/richtext	col -b %1 oder strings %1	bsdmainutils
text/sgml	col -b %1 oder strings %1	bsdmainutils
text/tab-separated-values	col -b %1 oder strings %1	bsdmainutils

Zum Extrahieren können verschiedene Werkzeuge genutzt werden. Viele Unix Sites benutzen "strings", welches versucht Text in Dateien zu erkennen - allerdings nicht mit der Genauigkeit eines speziellen Werkzeuges für den MIME Typ.

Stelle sicher, dass der Systembefehl seine Ausgabe auf dem Bildschirm (Standardausgabe) und nicht in eine Datei vornimmt. Probiere den Befehl vorher in einer Konsole und lies die Anleitung. Z.B. benötigt pdftotext ein nachfolgende "-".

Ggf. muss der Tiki Cache nach der Installation eines neuen Handlers geleert werden, damit das System den Handler erkennt.

Es ist vorteilhaft, Fileinfo zu installieren, um falsch indentifizierte MIME-Typen zu vermeiden.

Verwandt:

- <http://stosberg.net/odt2txt/>