

Sisis-Briefe

Letzte Aktualisierung Montag, 1. Oktober 2007

Das Bibliothekssystem Sisis Sunrise von OCLC Pica erzeugt seit ein paar Versionen die Briefe, die an Benutzer verschickt werden, in XML bzw PCL. Da Bibliotheksmitarbeiter die versandten Briefe im Nachhinein durchsuchen können müssen, tut eine Konvertierung not. Die beschriebene Methode setzt auf den Briefen in PCL sowie den im Klartext gespeicherten E-Mails auf, konvertiert diese und macht sie durchsuchbar.

Die Lösung

Diese Lösung geht davon aus, dass die PCL-Dateien bereits in PDF konvertiert worden sind. Das ist einfach und kostenfrei möglich mit dem Tool ghostpcl und einem dem folgenden ähnlichen Script:

```
if [ -f ${1} ]
then
PCLFONTSOURCE="/usr/local/ghostpcl_1.41p1/urwfonts"
NAME=`echo -n ${1} | sed 's/.pcl/.pdf/g`
/usr/local/bin/pcl6 -sDEVICE=pdfwrite -sOutputFile=${NAME} -dNOPAUSE ${1}
fi
```

Umbaut mit einem Befehl, der über *.pcl in Unterverzeichnissen von /var/spool/sisis/avserver/batch/alist/ läuft (oder, geschickter, direkt nach jedem Tageslauf die notwendigen Dateien direkt konvertiert), entstehen zu allen PCLs PDFs gleichen Namens. Dies wird hier nicht weiter beschrieben sondern vorausgesetzt.

Eine mögliche Lösung, die Briefe und E-Mails zu durchsuchen, ist nun, einen lokalen Spider aufzusetzen, der die PDFs und die Textdateien der E-Mails indiziert und eine Suchmaske bereitstellt. Das ist allerdings insofern ungünstig als dass sämtliche Briefe bzw E-Mails eines Tageslaufes in jeweils einer Datei landen. Das Suchergebnis ist also nicht viel wert, wenn es eine riesige PDF-Datei auswirft. Besser wäre, wenn jeder Brief in einer eigenen Datei landen würde, die man dann - sofern der Suchbegriff gefunden wird - direkt ausgeben oder verlinken kann. Die Aufgabenstellung lautet also: Splitten des PDFs in einzelne Dateien und Splitten der E-Mail-Sammeldatei in einzelne Textdateien.

Die Aufgabe, die PDFs zu splitten, kann mit dem Tool pdftohtml kostenfrei gelöst werden. Herunterladen, kompilieren und das entstandene pdftohtml nach /usr/local/bin/ kopieren. Dann funktioniert das Shellscript briefe4suche.sh (siehe unten).

Die anschließende Suche durchsucht einfach sequentiell alle im betreffenden Verzeichnis vorhandenen html- (die Briefe) und txt-Dateien (die E-Mails). Alle Fundstellen werden als Links ausgegeben. Es wird hier mit Absicht keine Datenbanklösung eingesetzt, um die Installation nicht zu verkomplizieren. Da die Suche so unter Umständen ein wenig dauern kann (hier geht es um sehr wenige Sekunden), wird die Ausgabepufferung im Programm deaktiviert, so dass man bereits Ergebnisse bekommt, während die Suche noch läuft.

Installation

- Einrichten eines Jobs, der mittels `ghostpcl` aus allen PCL-Dateien PDF-Dateien gleichen Namens (abgesehen von der Erweiterung, natürlich) erstellt. Dies als cronjob oder direkt im Tageslauf laufen lassen.

- Das Tool `pdftohtml` holen, kompilieren und nach `/usr/local/bin` kopieren.

- Ein Verzeichnis erstellen, in dem die Briefdaten für Apache erreichbar sein sollen. Bei mir ist das `/var/spool/isis/avserver/batch/alist/briefe4apache`

- Dieses Verzeichnis im Apache erreichbar machen. Dazu wird in die Datei `/usr/local/isis-pap/apache/conf/httpd.conf` geschrieben:

```
Alias /Briefe /var/spool/isis/avserver/batch/alist/briefe4apache
```

```
<Directory "/var/spool/isis/avserver/batch/alist/briefe4apache">
```

```
    AllowOverride AuthConfig
```

```
    Options +Indexes
```

```
</Directory>
```

Danach muss der Apache neu gestartet werden.

Es empfiehlt sich, den Zugriff auf dieses Verzeichnis auf Mitarbeiter zu beschränken. Dazu lege man eine geeignete `.htaccess` in das Verzeichnis (dafür ist das `AllowOverride`).

- Das Shellsript `briefe4suche` (s.u.) herunterladen, z.B. nach `/home/isis/sc` kopieren und ausführbar machen. Sofern die Pfade aus dieser Anleitung übernommen wurden, ist das Script direkt lauffähig. Ansonsten müssen die Pfade im Kopf des Scriptes entsprechend angepasst werden.

- Einen cronjob-Eintrag für den user `isis` für das Script definieren:

```
0 6 * * * /home/isis/sc/briefe4suche.sh >/dev/null 2>&1
```

Der Eintrag muss nach dem Tageslauf laufen. Einmal mehr schadet nicht, einmal konvertierte Dateien werden nicht erneut konvertiert.

- Das cgi-Script `suchbriefe.pl` (s.u.) nach `/usr/local/isis-pap/wwwdir/cgi-bin` kopieren (oder wo auch immer das `cgi-bin` des auf dem Sisis-Server laufenden Apache sich befindet).

- Das Suchformular `suchbriefe.html` (s.u.) auf einen Webserver der Wahl kopieren.

- In `suchbriefe.html` und `suchbriefe.pl` die Zeichenkette `meinsisis.server` durch die Adresse des eigenen Sisis-Servers ersetzen.

- Im Webbrowser die Adresse des Suchformulars ansurfen und testen. Sollte die Suche nicht funktionieren, die `/usr/local/isis-pap/apache/logs/error_log` (oder wo auch immer sie liegt) auf Fehlermeldungen prüfen.
Historie

- 9/07: Erste Version veröffentlicht

Bekannte Fehler/Probleme

- Die Liste der Suchergebnisse weist auf die Seiten in den Briefen, die die Treffer enthalten. Bei mehrseitigen Briefen sind dies u.U. die Folgeseiten, es wird in diesem Fall nicht der komplette Brief ausgegeben. Hinter dem Link auf die entsprechende Seite des Briefes wird daher ein Link auf die Sammelseite aller Briefe dieses Tages ausgegeben.

Alternativ kann man sich über eine manuelle Veränderung der URL der zweiten Seite (bspw .../Briefe/2007.09.28/BRF.EXT.28.09-98.html in /Briefe/2007.09.28/BRF.EXT.28.09-97.html) die Vorgängerseite holen. Abhilfe schaffen könnte eine Analyse der Seite daraufhin, ob es sich um eine Folgeseite handelt, und die Ausgabe der entsprechenden Links in der Seite selbst. Wenn jemand Lust hat ... ;-)

Download

Diese Programme sind kostenfrei im nichtkommerziellen Umfeld verwendbar. Das Programm darf ohne meine Einwilligung nicht kommerziell verwendet werden. Dies gilt insbesondere auch für Dienstleistungen, die die Installation oder Anpassung dieses Programmes beinhalten. Die Stellen, an denen mein Name steht, dürfen nicht verändert werden.

Selbstverständlich übernehme ich keine Haftung jeglicher Art für irgendwelche Schäden, die direkt oder indirekt durch die Nutzung dieses Programmes entstehen könnten.

Fehlerberichte und Anregungen zur Weiterentwicklung sind jederzeit willkommen.

Ich freue mich auch über eine Benachrichtigungsmail, wenn das Programm irgendwo eingesetzt wird.

- briefe4suche.sh : Vorbereiten der einzelnen Dateien, die durchsuchbar sein sollen
- Dasselbe, aber im Browser lesbar
- suchbriefe.pl : cgi-Perlscript, welches die Suche über die bereitgestellten Dateien ermöglicht
- Dasselbe, aber im Browser lesbar
- suchbriefe.html
- Dasselbe, aber als Sourcecode