Fazit

Beim Zugang zum CD-ROM-Service stellt das Web-Interface eine wesentliche Erleichterung in der Benutzung dar. Da alle WWW-Seiten dynamisch aus Konfigurationsdateien generiert werden, ist es schnell und einfach möglich, neu hinzu gekommene CD-ROMs in die Menüstruktur aufzunehmen.

Das Web-Interface wurde im Juni 2000 – kurz vor dem Erscheinen dieses Heftes – auf dem WWW-Server des RZ bereitgestellt. Es bleibt abzuwarten, wie der neue Zugang von den Nutzern des CD-ROM-Service angenommen wird.

Literatur:

1. Wendland, B.: Der CD-ROM-Service der Universitätsbibliothek. RZ-Mitteilungen Nr. 17 / Februar 1999 http://www.hu-berlin.de/rz/rzmit/rzinhalt.html#nr.17
2. Webseite des CD-ROM-Service: http://www.hu-berlin.de/rz/cd-rom-service

Bert Wendland
bwendland@rz.hu-berlin.de

ht://dig – Die neue Suchmaschine, Möglichkeiten der dezentralen Nutzung

Im World Wide Web (WWW) gibt es verschiedene Wege, Informationen zu finden. Eine häufig genutzte Möglichkeit ist die Volltextsuche. Derzeit gibt es knapp 30 Webserver an der HU, die von der neuen Suchmaschine indexiert werden.

Bis November 1999 kam auf dem HU-Webserver die Harvest-Suchmaschine zum Einsatz. Es gab mehrere Gründe, warum Harvest abgelöst wurde, einer war, dass diese Suchmaschine nicht mehr weiterentwickelt wurde. Nun wird ht://dig (*http://www.htdig.org/*) eingesetzt.

Hier eine Auswahl an Fähigkeiten von ht://dig:
- Exakte Wortsuche
- Fehlertolerante Wortsuche
- Boolean-Suche (logische Verknüpfung von Suchbegriffen)
- Wortstammsuche (verschiedene mögliche Endungen des Wortes werden gesucht, dabei kommen das ispell-Wörterbuch und das entsprechende Affix-File zum Einsatz)
- Synonyme
- Stopp-Wort-Liste (Liste von Wörtern, die nicht durchsucht werden sollen, z. B. WWW, der, die, das ...)

Es lassen sich individuelle Suchformulare für den eigenen Webserver oder die eigene Homepage erstellen. Folgendes HTML-Formular schränkt die Suche auf den Web-Server der HU ein und schliesst dabei die Homepage des Rechenzentrums aus:

```
<html>
...
<form action= "http://www.hu-berlin.de/
     cgi-bin/htsearch">
Suchbegriff:
<input type="text" name="words" size="30">
<input type="hidden" name="config"
      value="htdig">
<input type="hidden" name="restrict"
      value="www.hu-berlin.de">
<input type="hidden" name="exclude"
      value="/rz/">
</form>
...
</html>
```

Sinnvoll sind hier Suchformulare, die nur bestimmte Verzeichnisse einschließen, etwa für den Bestand der Studienabteilung oder separat für das Vorlesungsverzeichnis.
In der folgenden Tabelle werden die wichtigsten Formular-Elemente kurz erläutert. Alle weiteren Formular-Elemente befinden sich auf der Homepage von ht://dig (*http://www.htdig.org/hts_form.html*).

Formular-Element	Beschreibung
config	legt die Konfigurationsdatei der Suchmaschine fest. Bei individueller Konfiguration muss der entsprechende Name eingetragen werden, ansonsten ist "htdig" Standard.
words	dient der Eingabe der Suchbegriffe.
restrict	Der Inhalt (URL, Pfad etc.) dieses Formularelements schränkt die Suche auf den entsprechenden Wert ein.
exclude	dient zum Auschließen bestimmter URLs.

Die Ausgabe der Suchergebnisse lässt sich ebenfalls anpassen. Dazu ist es jedoch notwendig, die einzelnen HTML-Template-Dateien zu erstellen und dem Web-Administrator (*webadm@rz.hu-berlin.de*) zuzusenden, da diese Dateien auf dem WWW-Server der HU installiert werden müssen. Vom Web-Administrator wird eine entsprechende Konfigurationsdatei erzeugt, deren

Name in das Formular eingetragen werden muss. Hier eine kleine Übersicht, welche Templates unbedingt benötigt werden:

HTML-Template-Datei	Beschreibung
header.html long.html footer.html	Die Suchergebnisse werden aus header.html, long.html und footer.html in dieser Reihenfolge zusammengesetzt, wobei long.html für einen Sucheintrag genutzt wird und somit mehrfach hintereinander auftreten kann.
syntax.html	Wenn bei der Boolean-Suche die logischen Operatoren (and, nor, not) falsch verwendet werden, wird diese Seite ausgegeben.
nomatch.html	Wenn das Suchwort kein Ergebnis gebracht hat oder das Suchwort in der Stopp-Wort-Liste steht, wird diese Seite ausgegeben.

Dabei ist zu beachten, dass alle Referenzen (Links, Bild-Sourcen, Style-Sheet-Referenzen etc.) nicht als relative, sondern nur als absolute Adressen (z. B. „*http://www2.hu-berlin.de/fpm/*“ statt „../“) angegeben werden müssen. Die Templates selbst und eine Beschreibung der Ersetzungsmarkierungen findet man auf der Homepage von ht://dig (*http://www.htdig.org/hts_templates.html*).

In der Konfiguration von ht://dig kann festgelegt werden, wie die einzelnen Fundstellen gewichtet werden sollen. Diese Angaben haben Einfluss auf das Ranking der Fundstellen. Wird beispielsweise ein Suchbegriff in den Keywords eines HTML-Dokumentes gefunden, wird dieses Dokument an der Spitze der Auflistung angezeigt.

Die Standardeinstellungen sind folgende:
keywords_factor: 100 (4 Sterne)
title_factor: 100 (3 Sterne)
meta_description_factor: 50 (2 Sterne)
text_factor: 1 (1 Stern)
max_description_length: 60 (Ein Eintrag im Tag "description" wird nur mit 60 Zeichen berücksichtigt.)
max_descriptions: 5 (Es werden höchstens 5 "description"-Einträge berücksichtigt.)
maximum_word_length: 32

Für Web-Administratoren bietet sich so die Möglichkeit, gezielt zu steuern, welches Dokument gegenüber anderen zu einer Thematik zuerst angezeigt werden soll. Manchmal sollte man also auch auf einen Eintrag in den keywords verzichten, um einem anderen Dokument, das die Thematik vielleicht grundlegender behandelt, den Vorzug zu geben.

Und noch ein Hinweis: Die Suchmaschine erfasst nur solche Dokumente, die – beginnend bei *www.hu-berlin.de/index.htm*l – irgendwie über einen Link angekoppelt sind. Entsteht in einem Baum eine Lücke, weil die Dokumente nicht miteinander verlinkt sind, werden alle nachfolgenden Dateien nicht indexiert!

Daniel Rohde
d.rohde@rz.hu-berlin.de

Tools & Tipps

Link-Checker: Xenu's Link Sleuth (Xenu)

Jeder, der eine mindestens mittelgroße Web-Site verwalten muss, kennt das Problem der nicht mehr gültigen Links. Xenu kann diese Fehler zwar nicht beseitigen, aber sie wenigstens aufspüren. Durch seine einfache Installation und Handhabung kann es ein sehr nützliches Werkzeug für jeden Web-Administrator sein.
Xenu ist unter http://home.snafu.de/tilman/xenulink.html kostenlos erhältlich.

Die Windows-Software verfügt über folgende Features:
- einfaches Benutzer-Interface,
- ermöglicht die Eingrenzung des Suchvorgangs auf ein Unterverzeichnis,
- prüft Links innerhalb von HTML-Dateien, Image Maps, Java Applets, Skripten usw.,
- erstellt Reports in HTML-Form, die neben allen Fehlerausschriften auch eine Site-Map des untersuchten Bereiches beinhalten,
- prüft ungültige Links jederzeit erneut,
- unterstützt SSL (https://).