SELBST SUCHEN UND CRAWLEN MIT YACY

# Du bist Suchmaschine!

Wir schreiben das Jahr 2007: Die gesamte Netzgemeinde sucht mit Google. Nur eine kleine Gruppe Idealisten widersetzt sich dem Monopolisten und werkelt munter an einer eigenen freien Suchmaschine: YaCy, einer verteilten Engine, die auf normalen PCs als Proxy laufen kann und das Peer2Peer-Prinzip nutzt.

VON **MATTIAS SCHLENKER**

Alles Spinner? Mitnichten: Hinter YaCy steckt ein interessantes Konzept. Im Gegensatz zu gewöhnlichen Suchmaschinen, bei denen Bots ausschwärmen, um das Web zu indexieren, verwendet YaCy einen integrierten Proxy-Server als Quelle für den Index. Das schlägt gleich mehrere Fliegen mit einer Klappe: Zunächst einmal hält sich der Traffic in Grenzen, da indexierte Seiten sowieso von einem menschlichen Nutzer angefordert werden, also kein zusätzlicher Suchmaschinen-Traffic anfällt. Zudem dient es der Aktualität: Häufig besuchte Seiten werden bei jedem Aufruf aktualisiert, weniger häufig besuchte Seiten verschwinden irgendwann in der Bedeutungslosigkeit. Die letztgenannte Funktion ist übrigens ein gutes Indiz für die tatsächliche Beliebtheit einer Seite: Während Google und Co. primär vom Linkcount, also der Zahl der eingehenden Kanten, auf die Beliebtheit einer Seite schließen, zählt YaCy tatsächliche Besucher.

Damit nicht jeder Netzwerkknoten, auf dem YaCy läuft, für sich allein gestellt ist, tauschen die Knoten Teile des Indexes und Suchergebnisse aus. Die dafür verwendeten Algorithmen entsprechen den Flooding-Algorithmen von P2P-Anwendungen, weshalb YaCy auch als Peer-to-Peer-Suchmaschine bezeichnet wird. Aber nicht nur als indexierender Proxy lässt sich YaCy verwenden, es lassen sich auch ganz klassisch Crawls starten, für die ein „Bot“ losgeschickt wird. Auch dieser Funktionalität kommt die Peer2Peer-Struktur zugute: Je nach Konfiguration können für einen Crawl andere Knoten herangezogen werden, die gerade nichts zu tun haben - oder andersherum betrachtet: Man lässt seinen Knoten beim Crawlen mithelfen.

## Wieviel YaCy für welchen Einsatzbereich?

**› *Einzelarbeitsplatz***

**Wenn Sie 128 MByte RAM und wenigstens 1 GByte Festplattenplatz abzwacken können, lässt sich YaCy „nebenher“ auf jedem Einzelarbeitsplatz der Anderthalb-Gigahertz-Klasse verwenden. Mit dem richtigen Nice-Wert macht sich die recht CPU-intensive Indexierung kaum bemerkbar. Die Indexierungstiefe des Proxies sollte auf *0* stehen, damit nur die über den Proxy angeforderten Inhalte indexiert werden müssen. Gegen gelegentliche nächtliche Crawls spricht genauso wenig wie gegen gelegentliche Mitbenutzung durch einzelne Rechner im Netzwerk.**

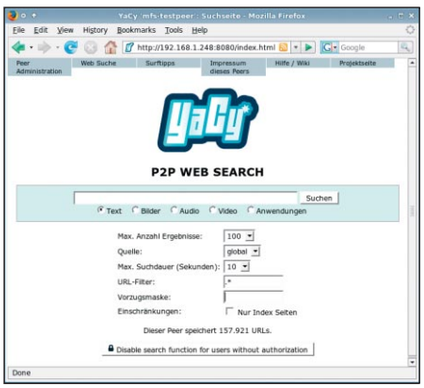
**› *Büro-/Heimnetz***

**Wer einen Datei- und Druckserver der Gigahertz-Klasse oder höher betreibt, kann diesen für YaCy mitverwenden. Im Büronetz des Autors (drei Arbeitsplätze) läuft YaCy auf einem älteren Athlon 2000 in einer Xen-Instanz mit 384 MByte RAM; genauso hoch ist das eingestellte Speicherlimit von YaCy. Dank einer Proxy-Indexierungstiefe von *1* erfasst YaCy auch alle direkt verlinkten Seiten.**

**Speicher- und CPU-Ausbau erlauben gelegentliche Crawls neben der Abarbeitung von Proxy-Indexierungsaufträgen. Wer wenigstens 512 MByte RAM und 2 GHz zur Verfügung hat, kann auch die *Pro*-Version von YaCy verwenden, die auch PDF und andere Formate indexiert. Mit einer Indexiertiefe von *0* lässt sich die Anzahl der möglichen Clients deutlich steigern, allerdings wird der Index nicht so schnell aufgebaut.**

Was gecached und indexiert wird, kann jeder YaCy-Administrator selbst bestimmen. So können „schwarze Listen“ von anderen Hosts importiert oder selbst erstellt werden, die für Proxy und Crawler gleichermaßen dienen. Mit diesen lassen sich beispielsweise allzu aggressive Werbeeinblendungen (solche, die ein sehr detailliertes Nutzerprofil erstellen), ganz ausblenden oder Schmuddelseiten vom Remote-Crawl ausschließen. Sowohl die global verwalteten Blocker-Funktionen als auch das Caching machen YaCy besonders für das Heimnetzwerk oder kleine (aber auch mittelgroße Firmennetze) interessant.

Dennoch ist der Einsatz auf einem Einzelrechner möglich und sinnvoll: Bereits wer ein paar Prozent CPU-Leistung, 64 bis 128 MByte vom Arbeitsspeicher und 1GByte Festplattenplatz für den Proxy abzwacken kann, hat die Möglichkeit am YaCy-Netz teilzunehmen. Der Rechner muss dafür nicht die ganze Zeit angeschaltet bleiben. Da der Index stückchenweise auf die anderen Knoten verteilt wird, bleibt er - wie daraus resultierende Suchergebnisse - verfügbar. Viel Potenzial bietet YaCy als Suchmaschine für das Intranet, allerdings sollte bei diesem Einsatzbereich darauf geachtet werden, dass weder entfernte Suchanfragen noch ein Upload des Indexes stattfindet.




**YaCy ist eine (fast) gewöhnliche Suchmaschine...**

## Begehrlichkeiten

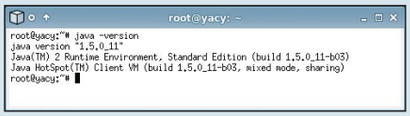
Sobald eine Technologie eine gewisse kritische Masse der Nutzung überschritten hat, wird sie für Spammer interessant. Diesen zweifelhaften Ruhm genießt mittlerweile auch YaCy. Prinzipbedingt wird der Spam-Anteil nie die Ausmaße wie bei E-Mail oder Blog-Spam erreichen, schließlich muss jeder Spammer zuerst einen YaCy-Knoten aufsetzen und ins YaCy-Netz integrieren. Dennoch tauchen immer wieder extrem tief gestartete Crawls über Pornoportale auf. Um eine einseitige Verschiebung des Indexes hin zu Schmuddelseiten zu verhindern, ist es deshalb sinnvoll, gelegentlich in die Remote-Crawl-Anfragen zu schauen und gegebenenfalls die „schwarze Liste“ um bestimmte URL-Muster zu aktualisieren. Da YaCy-Spam ein überschaubares Phänomen ist, sollte die Ein-

## Freiheitsliebend

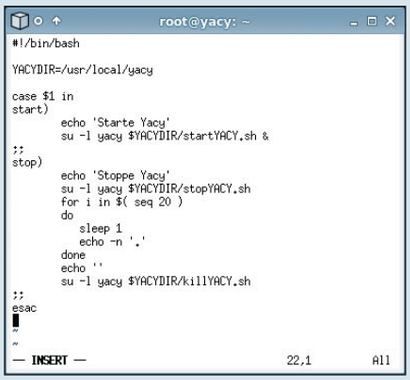
**Nicht nur in China und dem Iran gibt es Zensur im Internet. In den letzten Jahren kam es einige Male vor, dass Seiten sang- und klanglos aus dem Google-Index verschwanden oder per Sperrungsverfügung von der Düsseldorfer Bezirksregierung für Surfer aus Nordrhein-Westfalen unzugänglich gemacht wurden. Nur in seltenen Fällen, wie bei den verschwundenen Scientology-Dokumenten wird die Google-Zensur öffentlich diskutiert. Auch eine vollständige „Düsseldorfer Sperrliste“ ist nicht erhältlich - das Regierungspräsidium hat offensichtlich Angst davor, dass sich eine Debatte um die Sperrwürdigkeit der Seiteninhalte entwickelt.**

**Ein Ziel der YaCy-Entwickler war es, diese Formen von Zensur zu erschweren. Durch den dezentralen Index können nordrhein-westfälische YaCy-Knoten - selbst im hypothetischen Fall, dass deren Betreiber die Sperrliste erhalten - auf die Indexes bayrischer Knoten zurückgreifen. Und Surfer können zeitweise über einen YaCy-Knoten in einem anderen Staat oder Bundesland surfen, wenn eine Seite nicht erreichbar ist - sei es durch Routingprobleme oder eine Sperrungsverfügung. Dass die Düsseldorfer Bezirksregierung ihre Sperrungsverfügung an alle Knotenbetreiber in Deutschland verschickt, erscheint extrem unwahrscheinlich. Und selbst wenn, muss sie mit starkem Gegenwind rechnen, da die Inhalte der Liste bislang von keinem Gericht auf ihre tatsächliche Illegalität überprüft worden sind.**

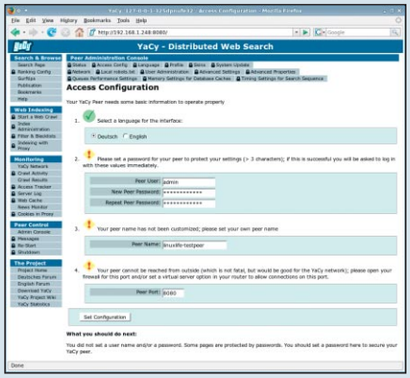
**Ein Freibrief für das vermeintlich anonyme Surfen über fremde Proxies oder die Indexierung rechtsextremer Propaganda ist das YaCy-Netz freilich nicht. Im Gegensatz zu Systemen wie *Tor* ist YaCy ein normaler Proxy, der mitloggt, von welchen Hosts Anfragen kamen. Beleidigende Kommentare in Foren, die über fremde Proxies getätigt werden, lassen sich so leicht zurückverfolgen. Die Angst vor einer Hausdurchsuchung, weil der YaCy-Crawler illegale Inhalte angefordert hat, ist äußerst gering: Offenbar können ermittelnde Behörden Aufrufstruktur und Referrer durchaus einem Bot zuordnen und wissen, dass Crawls nicht vom jeweiligen Knoten initiiert sein müssen. Zudem halten viele Knotenbetreiber Ausschau nach verdächtigen Crawls.**



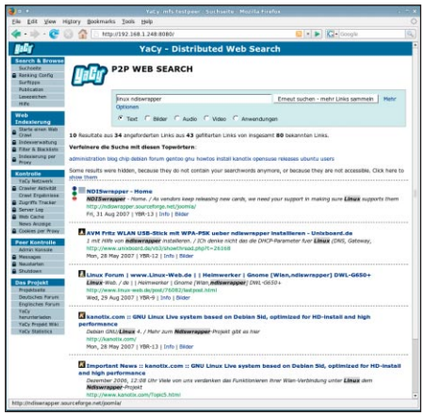

**YaCy setzt Suns Java in Version 1.4 oder 1.5 voraus, das inzwischen über die Repositories der meisten Distributionen nachinstalliert werden kann.**




**Damit YaCy automatisch beim Starten des Systems aufgerufen wird, sollten Sie ein RC-Script anlegen.**



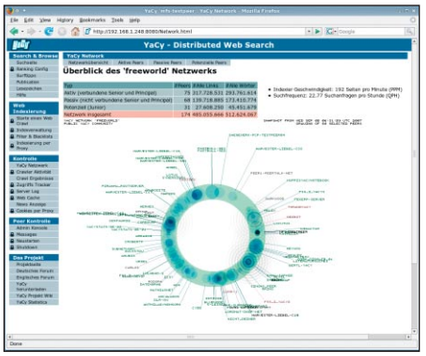

**Für die YaCy-Konfiguration wird ein Webfrontend mitgeliefert.**




**...die in vielen Bereichen bereits sehr gute Ergebnisse liefert.**

schränkung auf nicht allzu hohe Remote-Crawl-PPMs (Pages per Minute) und der tägliche Blick in die Crawl-Anfragen derzeit ausreichen. Mittelfristig müssen sich die YaCy-Entwickler jedoch Gedanken machen. So kann das Verhältnis zwischen nutzbarem Content und Spam beispielsweise dadurch in der Waage gehalten werden, dass man sich PPMs für Remote Crawls dadurch erkauft, dass man selbst zunächst Remote Crawls erlaubt. Auch eine übersichtlichere Integration fremder Blacklists sollte helfen.

Begehrlichkeiten positiver Natur weckt YaCy bei den Betreibern von Metasuchmaschinen wie *Clusty.com* oder *Metager.de*: Zumindest Metager zieht regelmäßig YaCy-Ergebnisse zur Verbesserung der Ergebnismenge heran. Prinzipiell ist auch eine Integration der YaCy-Ergebnisse in kommerzielle Suchmaschinen und Portale denkbar: Die von YaCy verwendeten Protokolle liegen offen, und die Suchmaschine selbst unterliegt der GPL. Ein gewisses Wohlverhalten („nehmen und zurückgeben“ oder Reputationseffekte) kann leicht von der YaCy-Community beeinflusst werden, indem sie allzu aggressiv anfragende Knoten nur bis zu einer gewissen Obergrenze mit Indexfragmenten und Suchergebnissen beliefert. Egal wie die Integration von YaCy-Suchergebnissen in größere Portale und Suchmaschinen aussehen wird: Sie gibt der großen Masse der Netznutzer eine bessere Einflussmöglichkeit als die recht passiven Indexierungsmechanismen, die derzeit vorherrschen, und macht die Vorteile von YaCy auf




**Tagsüber sind derzeit meist etwa 100 Senior Peers erreichbar.**

für Surfer zugänglich, die keinen eigenen Knoten aufsetzen wollen.

## Die Google-Konkurrenz?

Einige Online-Medien bejubeln YaCy bereits als Google-Konkurrenz, doch bis YaCy ein vollständiger Ersatz für die großen Suchmaschinen ist, werden noch einige Jahre vergehen. Als dieser Artikel entstand, umfasste das YaCy-Netz tagsüber meist um die hundert Senior-Peers, die etwa eine halbe Milliarde

## Installation von YaCy

**Obwohl YaCy als Tarball bezogen werden sollte, ist die Installation sehr geradlinig. Der indexierende Proxy setzt allerdings eine Java-Laufzeitumgebung voraus.**

**Führen Sie die folgenden Schritte als *root* durch – unter Ubuntu erhalten Sie mit *sudo su* – eine Rootshell:**

**1. Stellen Sie sicher, dass eine Sun-Java-Umgebung in Version 1.4 oder 1.5 installiert ist. Praktisch alle Distributionen bieten diese mittlerweile über ihre regulären Repositories an. Prüfen Sie mit dem Befehl**

```
java -version
```

**ob die richtige Version verlinkt ist.**

**2. Entpacken Sie den YaCy-Tarball in einem geeigneten Verzeichnis. Auf dem Testsystem pflegten wir eine BSD-ähnliche Ordnerstruktur:**

```
cd /usr/local
tar xvzf \
yacy_v0.54_20070802_4021.tar.gz
```

**3. Legen Sie einen Nutzer *yacy* an. Bei debianesken Distributionen gelingt dies auf der Kommandozeile mit:**

```
adduser yacy
```

**Passwort und weitere Nutzerdaten werden abgefragt, die Login-Shell können Sie auf */bin/false* oder */bin/nologin* setzen.**

**4. Übereignen Sie den YaCy-Ordner dem neuen Benutzer:**

```
chown -R yacy:yacy /usr/local/yacy
```

**5. Jetzt können Sie *yacy* von Hand starten, *su* sorgt dafür, dass dies mit den Rechten des Nutzers *yacy* passiert.**

```
su -l yacy \
/usr/local/yacy/startYACY.sh
```

**6. Damit YaCy bei jedem Hochfahren gestartet wird, empfiehlt sich ein RC-Script */etc/init.d/yacy*:**

```
#!/bin/bash

YACYDIR=/usr/local/yacy

case $1 in
start)
        echo 'Starte Yacy'
        su -l yacy
➥$YACYDIR/startYACY.sh &
;;
stop)
        echo 'Stoppe Yacy'
        su -l yacy
➥$YACYDIR/stopYACY.sh
        for i in $( seq 20 )
        do
           sleep 1
           echo -n '.'
        done
        echo ''
        su -l yacy
➥$YACYDIR/killYACY.sh
;;
esac
```

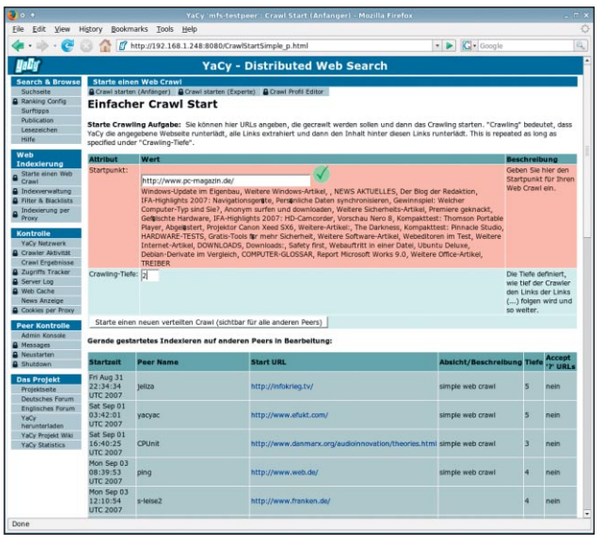
**Setzen Sie das Script auf ausführbar und verlinken Sie es in Ihrem Default-Runlevel (meist *2*, *3* oder *5*):**

```
chmod +x /etc/init.d/yacy
ln -s /etc/init.d/yacy \
/etc/rc2.d/S99yacy
```

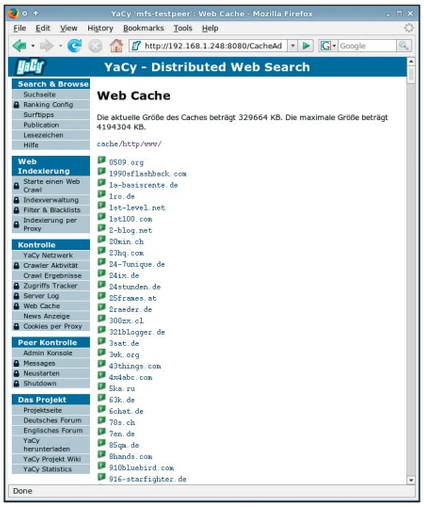
**7. Rufen Sie im Browser die YaCy-Konfigurationsseite auf – wenn YaCy auf dem selben Rechner läuft wie Ihr Browser mit http://localhost:8080, ansonsten über die IP-Adresse des externen Interfaces: http://192.168.1.248:8080**

**Setzen Sie den YaCy-Peernamen und vergessen Sie nicht, ein Administratorpasswort zu vergeben.**

**8. Soll Ihr YaCy-Knoten von außen erreichbar sein, öffnen Sie die Routerkonfiguration und leiten Sie Port 8080 auf Ihren YaCy-Host weiter.**




**Interessante Inhalte können als Crawl-Startpunkt definiert werden. Bei Crawl und Indexierung helfen andere Knoten mit.**




**Eine detaillierte Cache-Kontrolle erleichtert das Finden und ggf. Sperren unerwünschter Inhalte wie allzu aggressiver Werbung.**

Links und genauso viele Wörter indexierten. Was absolut betrachtet viel erscheint, ist nur ein Bruchteil dessen, was Google indexiert hat. Da sich der Großteil der YaCy-Knoten auf den deutschsprachigen Teil des Netzes konzentriert, liefert YaCy in einigen Bereichen bereits recht gute Suchergebnisse. Naturgemäß werden viele YaCy-Knoten von Verfechtern freier Software betrieben oder befinden sich als Zugangsproxy in Unternehmen, die sich der Softwareentwicklung oder dem Webdesign verschrieben haben. Dementsprechend hochwertig sind die Treffer bei Themen rund um Programmierung, freie Software, Webdesign - aber auch Netzwerksicherheit oder Datenschutz. Kulturelle Themen oder die schönen Künste hingegen sind subjektiv etwas im Hintertreffen, wobei einige in den letzten Monaten neu hinzugekommene Bibliotheksproxies in Universitäten aber für etwas mehr Ausgewogenheit sorgen. Bis Google und Co. ernsthaft und weltweit attackiert werden können muss die Gesamtzahl der Knoten aber in niedrige fünfstellige Bereiche vordringen, für eine sehr gute Indexierung deutschsprachiger Seiten sollte dagegen eine vierstellige Zahl ständig aktiver Knoten vollkommen ausreichen.

Deutlich komfortabler als bei den großen Suchmaschinen sind die Einflussmöglichkeiten des Nutzers auf die Suchergebnisse: YaCy bietet in Ansätzen Clustering, also die Gruppierung von Suchergebnissen nach Begriffen in kontextueller Nachbarschaft. Auf Deutsch: Wer nach „latex“ sucht, bekommt weitere Suchbegriffe vorgeschlagen, mit denen er die Such auf Themengebiete einschränken kann - beispielsweise „tex“ für das Textsatzsystem oder „matratze“ für ebensolche aus Naturkautschuk. Dazu bietet YaCy eine einfache Möglichkeit des User-Feedbacks: Der Surfer kann nützliche und unnütze Suchergebnisse voneinander unterscheiden und so Information, informative Werbung und Spam voneinander trennen.

Dass YaCy einige interessante Ansätze bietet hat auch Google erkannt: Dessen *Web Accelerator* bietet unter dem Vorwand eines schnelleren Surferlebnisses eine Kombination aus lokalem Proxy und der Verwendung des Google-Caches als Proxy. Die Indexierung erfolgt wie bei YaCy beinahe in Echtzeit. Wegen Sicherheitslücken in der Startphase verfolgt Google das Projekt derzeit nur auf Sparflamme, es besteht jedoch kein Zweifel daran, dass der Suchmaschinengigant dem Web Accelerator bei wachsender Konkurrenz mehr Ressourcen widmen wird.


**Die YaCy-Bar bietet Direktzugriff auf erweiterte Funktionen wie gemeinsam genutzte Bookmarks oder die schwarze Liste.**

## Weiterführende Links

**http://www.yacy.net**
**Homepage- und Download-Seite des YaCy-Projektes**

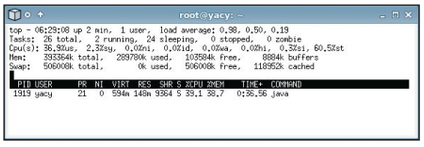
**http://yacy-websuche.de**
**Großes YaCy-Portal mit Wiki und Foren**

## Grenzen des Konzeptes

Wie bei vielen anderen P2P-Technologien ist ein Bottleneck, der YaCys Effizienz derzeit einschränkt, die stark asymetrische Bandbreitenverteilung zwischen Up- und Download. Fordert ein Knoten Suchergebnisse bei anderen Peers an, kann es einige Sekunden dauern, bis dieser seine Ergebnisse zurückliefert - insbesondere, wenn der entfernte Knoten andere Anfragen abzuarbeiten hat. Gegen Googles Cluster mit einem dedizierten Netzwerk und einem globalen Dateisystem ist in dieser Hinsicht kein Ankommen. Immerhin skaliert YaCy beinahe logarithmisch, wodurch sichergestellt ist, dass eine große Knotenzahl auch ausgenutzt werden kann. Soll YaCy mittelfristig Millionen von Clients bedienen können, wird allerdings kein Weg an einzelnen Superknoten - die als Cluster ausgeführt sein können - vorbeiführen. Diese können dann die Mehrzahl der „Brot- und Butter-Suchergebnisse“ in Sekundenbruchteilen liefern, während die vielen kleineren Peers hochwertige Detailsuchergenisse bereitstellen, wenn der Surfer einige Sekunden bereit ist zu warten.

## Fazit

Bis zu einem vollwertigen Google-Ersatz ist es noch eine Weile hin. Dennoch kann YaCy schon heute die großen Suchmaschinen sinnvoll ergänzen. Der Einsatz lohnt sich besonders, wenn sowieso ein Proxy für das lokale Netz gesucht wird und eine Filterfunktion für unerwünschte Inhalte gefragt ist. **jkn**




**Der Büro-Proxy mit Indexierungstiefe 1 erzeugt eine spürbare Systemlast beim Indexieren.**