

# Infos & Kommunikation

---

## Mit Harvest durchs Internet

*Oliver Göbel / Gabriele Mayer*

[Die Komponenten](#)

[Die Konfigurationen](#)

[Der Gatherer](#)

[Der Broker](#)

[Harvest Server Registry \(HSR\)](#)

[WWW Interface](#)

[Summary Object Interchange Format \(SOIF\)](#)

[Das Cache Subsystem](#)

[Der Replicator](#)

[Harvest an der Universität Stuttgart](#)

---

## Mit Harvest durchs Internet

*Oliver Göbel / Gabriele Mayer*

Harvest ist ein integriertes Paket von Werkzeugen, um im Internet Information von verteilten Systemen zu sammeln, extrahieren, organisieren, suchen, cachen und zu replizieren. Es ist ein frei verfügbares Stück Software, das mit moderatem Aufwand an Konfigurations- und Installationsarbeit Möglichkeiten zur komfortablen Verwaltung und Suche von und nach Information aus dem Internet bietet. Dies wird durch verschiedene Arten der Indizierung ermöglicht. Unter anderem bietet Harvest eine verteilte Indizierung an, die es erlaubt, die Netzlast, die durch einen Indizierungsvorgang entsteht, erheblich zu reduzieren. Im Rahmen des InfoWin-Projektes, an dem das Rechenzentrum maßgeblich beteiligt ist, soll Harvest auf verschiedenen WWW- und ftp-Servern in ganz Europa als Indizierungswerkzeug eingesetzt werden.

### Die Komponenten

Harvest setzt sich aus mehreren Subsystemen zusammen, die gleichzeitig in Betrieb sind. Der *Gatherer* sammelt Information zur Indizierung (z.B. Schlüsselwörter, Autorennamen, Titel usw.), die auf *Provider Sites* (ftp-, WWW-Servern, ...) verfügbar sind. Der *Broker* erhält diese Information von einem oder mehreren Gatherern, unterdrückt die Duplikation von Information, indiziert sie inkrementell und stellt eine WWW-basierte Benutzungsoberfläche für Suchanfragen zur Verfügung. Der *Replicator* repliziert Broker, die an anderen Stellen im Internet betrieben werden. Das *Object Cache Subsystem* optimiert den Zugriff auf häufig angefragte Informationen. Und schließlich gibt es noch den *Harvest Server Registry* (HSR), einen Broker, der an der University of Colorado in Boulder läuft und eine Datenbasis über alle im Internet installierten Broker, Gatherer, Replikatoren und Object

Cache-Systeme verwaltet. Am RUS wird bereits das Harvest Object Cache System als Proxy für die WWW-Server eingesetzt (s. [BI. 1+2/96](#), [S. 29 Neuer WWW-Proxy-Server](#)).

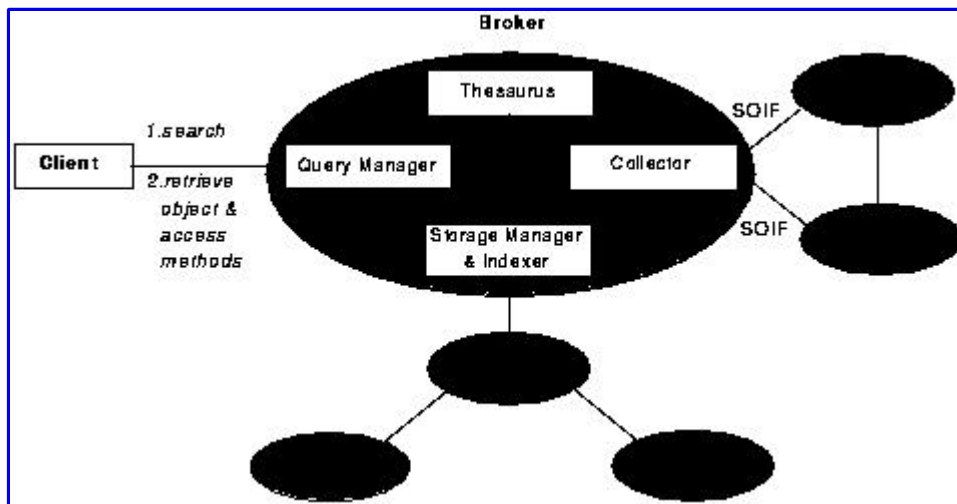


Abb. 1: Komponenten von Harvest

## Die Konfigurationen

Gatherer und Broker können auf unterschiedliche Weise konfiguriert werden (s. Abb. 2). Schon zur Indizierung von lokalen Daten können auf einem Rechner mehrere Gatherer einen oder mehrere Broker mit Information versorgen. Weiterhin können Gatherer, die verteilt im Netz betrieben werden, einen oder mehrere Broker, die auf der indizierenden Maschine laufen, bedienen (s. Abb. 2, rechte Konfiguration). Harvest kann jedoch auch Server indizieren, auf denen keine Harvest Gatherer laufen. Zu diesem Zweck holt ein lokal betriebener Gatherer mit Standard Retrieval-Protokollen, wie ftp, Gopher, HTTP oder NNTP, die Daten remote vom Provider (s. Abb. 2, links). Die stärkeren Linien verdeutlichen die höhere Server- und Netzbelastung, die dafür in Kauf genommen werden muß. Einen Gatherer lokal zu betreiben, ist zwar wesentlich effizienter (s. Abb. 2, rechts), doch ist ein nichtlokaler Gatherer besser als viele Server, die unabhängig voneinander Index-informationen sammeln, weil viele Broker oder andere Suchdienste den verteilt erzeugten Index mitbenutzen können.

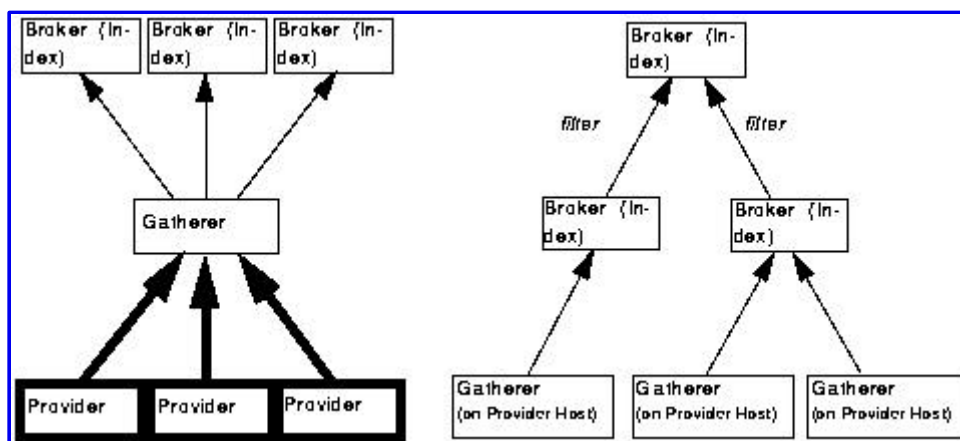


Abb. 2: Konfigurationsmöglichkeiten

Schließlich kann ein Broker so konfiguriert werden, nicht nur Gatherer abzufragen, sondern auch von anderen Brokern Daten zu beziehen, um einen Index aus weiträumig verteilten Informationen zu erstellen. Die Netzbelastung wird dadurch effektiv reduziert, da der Indizierungsprozeß verteilt abläuft.

Das Query Interface, über das der Broker die Information holt, läßt das Filtern oder Verfeinern der Information - und damit eine Erhöhung des Informationswertes durch Einarbeitung von Kontextinformation - zu.

## Der Gatherer

Wie bereits erwähnt, ermittelt der Gatherer mit Hilfe verschiedener Standard-Zugriffsmethoden, wie ftp, Gopher, HTTP, NNTP und lokale Dateien, Information und erzeugt aus den typspezifisch zusammengefaßten Daten strukturierte Indexinformation. Beispielsweise kann ein Gatherer aus einem ftp-Archiv einen technischen Bericht holen und Autor, Titel sowie Abstract aus dem Dokument extrahieren. Harvest Broker oder andere Suchdienste holen dann die Indexinformation vom Gatherer und stellen sie in einem durchsuchbaren Index vermittels eines WWW-Interfaces zur Verfügung.

Der Gatherer setzt sich aus verschiedenen Komponenten zusammen. Das Programm *Gatherer* steuert den durchgehenden Prozeß der Aufreihung und Zusammenfassung von Datenobjekten. Die strukturierte Indexinformation, die der Gatherer sammelt, ist als Liste von Attribut-Wertpaaren im *Summary Object Interchange Format* (SOIF) dargestellt. Der Gatherer-Dämon *Gatherd* stellt Brokern die Gatherer-Datenbank zur Verfügung; er läuft nach einer abgeschlossenen Gathering-Sitzung im Hintergrund. Ein stand-alone Gatherer-Programm dient dem Gatherd-Dämon als Client. Es kann von der Kommandozeile aus für Tests benutzt werden und wird vom Broker verwendet. Der Gatherer speichert gefundene Objekte auf einem lokalen Platten-Cache, der nicht vom Object Cache Subsystem verwaltet wird.

## Der Broker

Der Harvest Broker besitzt eine flexible Indizierungsschnittstelle, die verschiedene Search Engines wie zum Beispiel WAIS oder Nebula zur Indizierung und Anfrage an den Index einsetzen kann. Als Standard Search Engine wird *Glimpse* genutzt.

Glimpse unterstützt bei der Formulierung von Suchanfragen folgende Eigenschaften:

- case-sensitive und case-insensitive Anfragen
- Patternmatching: exakt, Teilwörter oder Ausdrücke mit mehreren Wörtern
- Bool'sche Kombinationen von Wörtern (mit AND/OR)
- approximatives Patternmatching, Schreibfehler können erlaubt sein
- strukturierte Anfragen, Treffer können an bestimmte Attribute gebunden werden
- Darstellung von einzelnen Zeilen oder ganzen Datensätzen, in denen ein Treffer vorkam, beispielsweise Zitate
- Begrenzung der Trefferanzahl
- vereinfachte reguläre Ausdrücke (Wild Cards).

Um die verschiedenen Optionen einzustellen sind im WWW Interface Buttons vorgesehen, die eine leichte Bedienung ermöglichen.

## Harvest Server Registry (HSR)

*Harvest Server Registry* ist ein besonderer Broker, der von der University of Colorado, Boulder, betrieben wird, und eine Datenbasis über Harvest Gatherer, Broker, Cache und Replicator im Internet bereithält. Ist ein Harvest Gatherer, Broker oder Cache erfolgreich eingerichtet, sollten die Server im HSR mit einer vorhandenen Registrierungsseite eingetragen werden.

## WWW Interface

Zum Query Manager und zur Administration des Brokers wurde ein World Wide Web Interface implementiert. So wird gewährleistet, daß mit Web-Browsern leicht auf den Broker zugegriffen werden kann und eine systemunabhängige Benutzungsschnittstelle zur Verfügung steht. Dieses WWW Interface umfaßt einige HTML Files und ein paar Programme, die das *Common Gateway Interface* (CGI) verwenden.

Im einzelnen besteht es aus:

- HTML-Dateien, die Forms für eine graphische Benutzungsoberfläche (GUI) verwenden
- CGI-Programmen als Gateway zwischen Benutzer und Broker
- Hilfedateien für den Benutzer.

## Summary Object Interchange Format (SOIF)

Harvest Gatherer und Broker kommunizieren mit Hilfe eines Attribute Value Stream-Protokolls, dem *Summary Object Interchange Format* (SOIF). Gatherer erzeugen Inhaltszusammenfassungen für individuelle Objekte in SOIF und geben diese Zusammenfassungen an Broker weiter, die sie sammeln und indizieren wollen. SOIF bietet Möglichkeiten Objektsammlungen zu klammern, damit Harvest Broker SOIF-Inhaltszusammenfassungen von einem Gatherer für viele Objekte in einem einzigen effizient komprimierten Stream holen können. Harvest Broker unterstützen Anfragen auf SOIF-Daten mit strukturierten Attribut-Wert-Anfragen und viele andere Anfragetypen.

## Das Cache Subsystem

Der Object Cache erlaubt Benutzern, ftp, Gopher und HTTP Daten schnell und effizient zu holen, ohne das Internet zu belasten. Der Harvest Cache ist mehr als eine Größenordnung schneller als der CERN Cache und andere bekannte Internet Caching-Systeme. Dies wird dadurch erreicht, daß für WWW- und Gopher-Zugriffe nicht geforkt wird, die Ein-/Ausgabe sowie DNS Lookups nichtblockierend erfolgen, Meta-Information, oft gefragte Objekte und DNS Lookups im RAM gecached werden. Der Cache kann in zwei verschiedenen Modi betrieben werden:

1. als Proxy Object Cache und
2. als httpd Accelerator.

## Der Replicator

Der Harvest Replicator verteilt Kopien einer Broker-Datenbank zu Repliken, die verteilt im Internet laufen. Replikation reduziert die Server-Belastung eines Brokers durch Lastverteilung und verbessert die Anfrage-Performance und Verfügbarkeit. Der Replicator versucht Netzwerkverkehr und

Server-Auslastung beim Verbreiten von Updates niedrig zu halten.

Der Replicator verwaltet einen Verzeichnisbaum aus Dateien. Ein Server muß als *Master Copy* ausgezeichnet sein. Updates davon werden auf alle anderen Repliken verbreitet und die Master Copy überschreibt jede lokale Änderung, die an individuellen Repliken gemacht wurde. Es ist möglich, eine replizierte Kollektion so zu konfigurieren, daß eine andere Master Copy getrennte Unterbäume verwaltet, um die Verantwortung zu verteilen. Jede replizierte Kollektion wird durch eine einzelne oder hierarchisch geschachtelte *Replication Group* exportiert, für die es eine Access Control List gibt. Wenn eine Replik zu einer Replication Group hinzugefügt wird, beginnt sie sich mit Daten zu füllen. Wächst eine Replication Group auf hunderte oder tausende von Repliken an, kann eine neue Gruppe erzeugt werden, um die Verwaltung zu erleichtern.

Dieser kurze Überblick über Harvest ist dem Harvest User`s Manual von Darren R. Hardy, Michael F. Schwartz und Duane Wessels von der University of Colorado, Boulder, entnommen.

Die Harvest Home Page ist unter der URL <http://harvest.cs.colorado.edu/> zu finden.

## Harvest an der Universität Stuttgart

Wie bereits erwähnt, soll Harvest im InfoWin-Projekt eingesetzt werden. Nach erfolgter Installation und erfolgreichem Einsatz im Rahmen dieses Projektes soll Harvest dann auch kurzfristig zur Indizierung der WWW- und ftp-Server der Universität Stuttgart eingesetzt werden. Hierzu wird ein weiterer Artikel in der BI. publiziert, der unter anderem auch empfohlene Konfigurationen für potentielle Teilnehmer enthalten wird.

Oliver Göbel, NA-5963

E-Mail: [Oliver.Goebel@rus.uni-stuttgart.de](mailto:Oliver.Goebel@rus.uni-stuttgart.de)

Gabriele Mayer, NA-5934

E-Mail: [Gabriele.Mayer@rus.uni-stuttgart.de](mailto:Gabriele.Mayer@rus.uni-stuttgart.de)