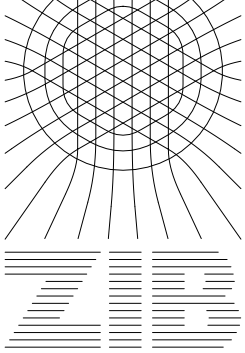
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Heilbronner Str. 10, D-10711 Berlin - Wilmersdorf



Judith Plümer Roland Schwänzl
Fachbereich Mathematik
Universität Osnabrück

Harvesting Mathematics

Ein Werkzeug zur Strukturierung merkmalsdefinierter Teilbereiche des WWW am Beispiel des elektronischen mathematischen Informationsangebots in der Bundesrepublik Deutschland

Diese Arbeit entstand im Rahmen des vom ZIB betreuten Vorhabens der Deutschen Mathematiker Vereinigung "Verbesserung des benutzerorientierten Zugriffs auf fachspezifische Online-Datenbanken und CD-ROM für Mathematische Institute der Bundesrepublik Deutschland". Das Vorhaben wurde vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie gefördert.

Technical Report TR 96–1 (Januar 1996)

Abstract: Nahezu flächendeckend sind die mathematischen Fachbereiche der BRD zum Jahresende '95 im WWW vertreten. Durch die relativ hohe Zahl von Servern entstehen Schwierigkeiten bei der Sichtung der angebotenen Information. Wir besprechen "Harvest" als brauchbares und gebrauchsfähiges Hilfsmittel zur Dokumentation.

Inhaltsverzeichnis

1 Einleitung

WWW Server sind Publikationsmedien. Sie sind Medien neuer Art, als technische Innovation vergleichbar der Einführung des Buchdrucks. Der Buchdruck ermöglichte erstmals die Erstellung einer großen Zahl von Kopien eines Werkes in kurzer Zeit. Erforderlich blieb aber ein beträchtlicher organisatorischer und technischer Aufwand zur Vervielfältigung und Verteilung, der vom Autor nicht geleistet werden konnte.

Dieser Umstand wurde in der Mathematik zugleich zur Qualitätskontrolle benutzt.

Der Einsatz von WWW Servern schafft gründlich Veränderungen:

1. Die Vervielfältigung eines mathematischen Texts ist trivial geworden. Eine notwendige Vorstufe war der Siegeszug von TeX als **die** Textformatierung für Mathematik.

2. Die Verteilung scheint mit dem Aufladen auf den Server erledigt.

3. Information der Öffentlichkeit erfolgt unmittelbar.

4. Qualitätskontrolle wird zeitlich hinter die Veröffentlichung gelegt.

5. Mathematiker werden mit Rechtsfragen konfrontiert, die - viele - Juristen nicht beantworten.

In dieser Arbeit geht es um den scheinbar erledigten Punkt 2.

Typischerweise präsentieren sich nicht einzelne Mathematiker mit eigenen Servern auf dem Internet, sondern es tritt der Fachbereich, das Institut oder der Sonderforschungsbereich als Organisator auf. Dabei verfolgt die jeweilige Organisation auch eigene Interessen und Aufgaben. Sie teilt der Öffentlichkeit die an ihr ablaufende Lehr- und Forschungstätigkeit mit.

Nimmt die Öffentlichkeit das Schaufenster und die darin ausgestellten Waren wahr?

Mit der Installation eines Servers und seiner Bestückung alleine ist das Verteilungsproblem noch nicht gelöst. Zunächst wird nur eine Stammkundschaft den neuen Laden betreten. Das kann nicht im Interesse der Autoren liegen. Erforderlich ist ein zugänglicher Dokumentardienst, der die lokalen Aktivitäten als Gesamtheit öffnet. Dafür sind Server übergreifende Indexierungswerkzeuge erforderlich, die inhaltliche Suche ermöglichen. Ein hierzulande von mathematischen Fachbereichen noch selten eingesetztes Werkzeug ist "Harvest". Uns bekannt sind nur noch Installationen in Kaiserslautern [5] und Rostock [6]. Es geht uns darum zu zeigen, daß es diese Software erlaubt unter Wahrung der Gestaltungvorstellungen der Betreiber einzelner Server, Gesamtaktivität darzustellen und damit zugleich den Anstrengungen der Fachbereiche und Institute eine bessere Resonanz zu verschaffen.

Wie heutzutage jede "anständige" Software, so hat auch Harvest seine kleinen und großen Bugs. Wir haben sicherlich nicht alle herausgefunden. Auf einige gehen wir im Detail ein.

Wir skizzieren grob den Aufbau von Harvest. Wer an den Aufbau eines eigenen Harv est Standorts denkt, sollte das ungewöhnlich gut lesbare User Manual [2] zu Rate ziehen. Eine reale Beispielkonfiguration wird in ihrer Funktionalität besprochen. Neben der Darstellung des technischen "Ist" werden auch einige "Solls" formuliert, von denen wir uns ohne wesentlichen Zusatzaufwand auf Serverseite eine Verbesserung des Gesamtwerts versprechen.

Es handelt sich dabei im wesentlichen um Konfigurationsdetails von WWW Servern, die bei Wunsch des Serverbetreibers, dem menschlichen Auge verborgen bleiben können. Wir bevorzugen PostScript als Format zur Distribution mathematischer Texte. Wir geben Hinweise auf den Nutzen von Indexfiles und die derzeit selten benutzten Vorteile der "META"-tags in HTML.

2 Zweckbestimmung mathematischer WWW-Server

Zur täglichen Arbeit eines Hochschulmathematikers gehört vor allem Lehre und Forschung. Zu beidem verfaßt er mehr oder weniger gute (viele) Texte. Einige Mathematiker

betätigen sich als Softwareproduzenten, sei es als Mitarbeiter bei der Erstellung und Entwicklung von Computeralgebraprogrammen oder Zusatzpaketen zu existierenden Systemen dieser Art, bzw. als Entwickler von Algorithmen in der angewandten Mathematik. Schließlich ist bei dieser Gelegenheit ausdrücklich Visualisierung und Simulation zu nennen. Zusammengefaßt: am Ende einer Anstrengung stehen TEXTE und Programme.

Ein mathematischer Fachbereich organisiert neben dem wöchentlichen Tee zum Kolloquium die Lehre und stellt die Infrastruktur zum Forschungbetrieb sicher. Dazu gehört heutzutage auch Netz- und Serverbetrieb. Der Fachbereich organisiert die Anwerbung von Studenten (schon zur Abdeckung seines Lehrdeputats). Kommt noch die "Studienberatung" für eingefangene Studenten hinzu und die Veranstaltung vieler Sitzungen, von Räten und Ausschüssen, Unterausschüssen und Kommitees, die für uns an dieser Stelle alle ohne Belang sind.

Eine einzige Ausnahme hiervon: Ein Endprodukt der vielen Sitzungen sind Prüfungs- und Studienordnungen. Ein für Studenten und auch für Lehrende relevanter Lesestoff. Damit ist im wesentlichen erfaßt, welche Art von Texten an einem Fachbereich produziert werden. Es fehlt noch eine Liste der Mitarbeiter, möglicherweise mit einer Beschreibung ihrer mathematischen Interessen. Ein Werbeprospekt wird auch über Substrukturen am Fachbereich, Forschungsgruppen und Einbindung in andere Organisationen berichten. Aus diesem Material schöpft ein Fachbereich sein eigenproduziertes Angebot auf seinem WWW-Server.

Für die eigenen Mitarbeiter und Studenten, ist der Web-Server das Tor zum mathematischen Material auf dem Internet. Der Webzugang soll es ermöglichen Preprint(sammlungen), elektronische Zeitschriften, Softwaresammlungen, Anschriften von Mathematikern und Organisationen schnell und zuverlässig auffinden zu können. Der Web-Server ist das Tor zum - und soll selbst Bestandteil des von [3] beschriebenen Math-Net sein.

Vom Web-Browser sollte Zugang zu Bibliotheken und Dokumentenlieferdiensten bestehen. Verlagsprospekte lassen sich heute schon einsehen. Spezialbuchhandlungen akzeptieren elektronische Orders. (Der Zugang zu den Online Datenbanken des Zentralblatts für Mathematik und ihrer Grenzgebiete, der Mathematical Reviews könnte ebenfalls noch in das Werkzeug integriert werden.)

Die Sorge für das schnelle und zuverlässige Auffinden von Informationen bedeutet für mathematische Autoren de facto die Übernahme verlegerischer Funktionen durch Fachbereiche und das derzeit informell sich entwickelnde Math-Net. Elektronische Zeitschriften - unabhängig von der sie tragenden Organisation - benötigen zumindest die Erwähnung als Link auf möglichst vielen Servern, um Leser zu gewinnen. Besser für die Reputation elektronischer Zeitschriften sind möglichst gutplatzierte Spiegel und Archive. Nur so können sie wirkliche Marktpräsenz erreichen. Dies gilt insbesondere auch für von kommerziellen Verlagen zu vertreibende elektronische Zeitschriften.

3 Zustandsbeschreibung

Von der im vorigen Paragraphen angegebenen Zweckbestimmung sind (fast) alle in der Bundesrepublik betriebenen Web-Server mehr oder minder weit entfernt. Das Bewußt-

sein, daß eine klare Struktur und die Konzentration auf das Wesentliche einen Wert an sich darstellen, das jede gute mathematische Arbeit durchzieht, muß noch auf den Aufbau von Web-Servern übertragen werden. Gebührt der Architektur des Institutsgebäudes wirklich der erste Platz auf der Titelseite? Schwer wiegt die an manchen Orten betriebene unkoordinierte Installation von Servern. Es ist definitiv eine verkehrte Idee, vom Server von Lehrstuhl "XY-Theorie I" zum Server des Lehrstuhls "XY-Theorie II" des gleichen Fachbereichs nur über eine Deutschlandkarte gelangen zu können. Unnützer Aufwand für die Besucher und damit verbunden Verkehrung des verlegerischen Elements ins Gegenteil. Bei größeren Fachbereichen ist es unter Umständen sinnvoll, mehrere Server zu betreiben und so Teilaufgaben zu delegieren. Die Einbettung in eine Verweisstruktur ist jedoch unerläßlich. Zu leicht wird sonst eine vielleicht wichtige Arbeit schlichtweg übersehen.

4 Entwicklungslinien

Die Autoren von [3] zielen auf ein dezentrales, verteiltes System für ein Math-Net ab. Verschärfend ist zu sagen, daß bei dem ausgeprägten Individualismus des Mathematikers ein zentralisiertes System gar nicht möglich wäre. Dies nicht auf Bundesebene, erst recht nicht international. Unter 3. haben wir Auswüchse auf Institutsebene kritisiert. Wie wird bei dieser Lage Informationsfluß und Kommunikation möglich? Wie es immer geschieht: Durch Aggregation von Verweisen an Knoten, die für diese Tätigkeit bekannt sind, wozu relative Konstanz gehört. Redundanz ist dabei nicht schädlich, im Gegenteil - sie macht ein Kommunikationsnetz erst tragfähig, weil im wesentlichen unverletzlich.

Wie wir sehen werden, leistet ein Programmsystem - wie beispielsweise "Harvest" - die notwendige Aggregation der Verweisstruktur.

Eines darf aber nicht übersehen werden: Forschritt in der Mathematik basiert nicht nur auf neuen Ideen und deren Mitteilung, sondern ganz wesentlich auf dem Erlernen, Verstehen und (kreativen) Anwenden von Vorhandenem und Geprüftem.

Prüfung (Referierungsprozess), Fixierung des Zustandes ("Publikation" im klassischen Sinn), Reviewing, Archivierung (Bibliotheken): Diese Stufen der Entwicklung von Mathematik sind nicht Gegenstand der hier in Frage stehenden Art von Dokumentation. Sie sind unverzichtbar und zumindest der Punkt "Prüfung" ist in keiner Weise automatisierbar. Ein Fachbereich kann die Prüfung nicht übernehmen, vielleicht die Archivierung.

5 Aufbau von Harvest

Harvest ist ein Werkzeug zum gezielten Sammeln von Informationen, die im Internet verteilt vorhanden sind, und deren Aufbereitung in einer Datenbank.

Harvest wurde an den Universitäten: University of Colorado Boulder, University of Arizona und University of South California von der IRTF-RD entwickelt. Derzeit ist unter *http://harvest.cs.colorado.edu/harvest/gettingsoftware.html* die Version 1.4 Patchlevel 1 zugänglich. Die Software wird in Karlsruhe gespiegelt (*ftp://ftp.ask.uni-karlsruhe.de/pub/infosystems/harvest/*). Harvest wird als Source distributiert.

5.1 Harvest

Harvest ist eine Kollektion von Dienstprogrammen, die teilweise bereits in der public-domain, bzw. dem Bereich freier Software existieren, deren Interaktion von Perl-Skripten gesteuert wird. Der lokal triviale Aufbau erlaubt den Einbau von Verbesserungen durch den jeweiligen Verwalter einer Installation.

5.2 Komponenten

Harvest besteht aus weitgehend unabhängig voneinander arbeitenden Hauptkomponenten:

- dem Gatherer
- dem Broker
- (dem Cache)
- dem Replicator.

Der Replicator ist ein Mittel zur Bildung von Netzen von Harvestinstallationen. Der Cache dient dem Performancetuning, er wird mittlerweile separat distributiert und ist nicht notwendiger Bestandteil einer Harvestinstallation.

5.3 Funktionsweise

Harvest ist auf das Vorhandensein eines http-Servers und für den Replicator auch auf einen funktionsfähigen anonymous ftp-Server angewiesen. Als weitere Softwarevoraussetzungen benötigt Harvest Perl (ab Version 4) und Gnu-Tools neueren Datums.

Der Gatherer

Der Gatherer erfüllt zwei Aufgaben. Das eigentliche Sammeln der Information und das Aufbereiten der Information für den Broker. Im weiteren bezeichnen wir mit Gathern das Sammeln und mit Essence das Extrahieren von Information. Gathern von Information erfolgt durch die Beauftragung des Servers Kontakt aufzunehmen, mit manuell eingetragenen http-, ftp-, gopher- und Hyper-G-Servern. Im letzteren Fall erfolgt die Kontaktaufnahme über das WWW-Gateway. **Die Konfigurierbarkeit des Gather-Prozesses ist eine der zentralen Ideen von Harvest.** Für jeden zu besuchenden Server können die Zugangsmethode, die Suchbreite und -tiefe gewählt werden. Individuell können Filter eingebaut werden (siehe 5.5). Die gefundenen Files werden zunächst in vollem Umfang übertragen. Mathematische Server sollten zur Reduktion der Netzlast grundsätzlich .dvi und .ps files in komprimierter Form anbieten. Harvest ist in der Lage diese vor der lokalen Weiterverarbeitung zu dekomprimieren. Nach Übertragung und gegebenenfalls notwendiger Dekomprimierung setzt die Auswertung der Files ein. Harvest nutzt wesentlich die durch die Endung vorgegebenen Heuristiken um bei der

Auswertung zu verzweigen. Ergebnis der Auswertung sind je nach Filetyp ASCII-Dateien mit mehr oder weniger strukturierter Zusatzinformation (Attributes) im Harvest eigenen SOIF (Summery Object Interchange Format). Die Standarddistribution sieht die folgenden Attribute vor (Harvest: user-manual [2], p. 75).

Attribute	Descirption
Abstract	Brief abstract about the object.
Author	Author(s) of the object.
Description	Brief description about the object.
File-Size	Number of bytes in the object.
Full-Text	Entire contents of the object.
Gatherer-Host	Host on which the Gatherer ran to extract information from the object.
Gatherer-Name	Name of the Gatherer that extracted information from the object. (eg. Full-Text, Selected-Text, or Terse).
Gatherer-Port	Port number on the Gatherer-Host that serves the Gatherer's information.
Gatherer-Version	Version number of the Gatherer.
Update-Time	The time that Gatherer updated the content summary for the object.
Keywords	Searchable keywords extracted from the object.
Last-Modification-Time	The time that the object was last modified.
MD5	MD5 16-byte checksum of the object.
Refresh-Rate	The number of seconds after Update-Time when the summary object is to be re-generated. Defaults to 1 month.
Time-to-Live	The number of seconds after Update-Time when the summary object is no longer valid. Defaults to 6 months.
Title	Title of the object.
Type	The object's type. Some example types are: Archive, Audio, Awk, Backup, Binary, C, CHeader, Command, Compressed, CompressedTar, Configuration, Data, Directory, DotFile, Dvi, FAQ, FYI, Font, FormattedText, GDBM, GNUCompressed, GNUCompressedTar, HTML, Image, Internet-Draft, MacCompressed, Mail, Makefile, ManPage, Object, OtherCode, PCCompressed, Patch, Perl, PostScript, RCS, README, RFC, SCCS, ShellArchive, Tar, Tcl, Tex, Text, Troff, Uuencoded and WaisSource
Update-Time	The time that the summary object was last updated. REQUIRED field, no default.
URL-References	Any URL references present within HTML objects.

Es ist für einen Harvest-Server Betreiber möglich, weitere (oder weniger) Attribute zu

definieren und nach Filetypen zu variieren. Grundsätzlich gilt natürlich, **daß nicht vorhandene Information auch nicht erzeugt werden kann**. So kann etwa aus einem PostScript-File das Attribut "author" nicht gewonnen werden. Auf die Auswahl der Keywords besteht Einflußmöglichkeit. Auf eine interessante Möglichkeit HTML Dokumenten "unsichtbare" verwertbare Information hinzuzufügen wird im manual hingewiesen (der HTML-Tag META!, siehe 8.).

.html Files:
HTML-Files werden zunächst unter sgmls geparsed. Fehlermeldungen des Parsers werden unterdrückt. Die als **korrekt** erkannten HTML-Tag's werden nach folgendem Schema in SOIF Attribute übersetzt (Harvest: user-manual [2], p. 19):

HTML ELEMENT	SOIF ATTRIBUTES
<A>	keywords,parent
<A:HREF>	url-references
<ADDRESS>	address
<B>	keywords,parent
<BODY>	body
<CITE>	references
<CODE>	ignore
<EM>	keywords,parent
<H1>	headings
<H2>	headings
<H3>	headings
<H4>	headings
<H5>	headings
<H6>	headings
<HEAD>	head
<I>	keywords,parent
<IMG:SRC>	images
<META:CONTENT>	$NAME
<STRONG>	keywords,parent
<TITLE>	title
<TT>	keywords,parent
<UL>	keywords,parent

Der in der Harvestdistribution mitgelieferte HTML Extraktor erzeugt aus:

Algebra I

Roland Schwänzl

Osnabrücker Schriften zur Mathematik, EHeft 1, 1993

Einführung in die Algebra II

Heinz Spindler

Osnabrücker Schriften zur Mathematik, EHeft 3, 1994

Algebraische Geometrie I

Heinz Spindler

Osnabrücker Schriften zur Mathematik, EHeft 5, 1995

Topologie I

Rainer Vogt

Osnabrücker Schriften zur Mathematik, EHeft 4, 1995

Übung Mathematik für Wirtschaftswissenschaftler, Aufgabensammlung

Roland Schwänzl

Osnabrücker Schriften zur Mathematik, EHeft 2, 1993

[Fig.1]

das Summary:

SOIF Object for: http://esther.mathematik.uni-osnabrueck.de/script/osm.html

---

```
@FILE { http://esther.mathematik.uni-osnabrueck.de/script/osm.html
update-time{9}: 811688551
description{8}: Scripten
last-modification-time{9}:      811688201
time-to-live{7}:        2419200
refresh-rate{6}:        604800
gatherer-name{4}:       osna
gatherer-host{35}:      esther.mathematik.Uni-Osnabrueck.DE
gatherer-version{3}:    1.0
type{4}:        HTML
file-size{4}:   1215
md5{32}:        2f777627e56cf1cc1a14aec27c5d02f9
body{506}:       Algebra I Roland Schw\344nzl Osnabr\374cker
Schriften zur Mathematik, EHeft 1, 1993 Einf\374hrung in die
Algebra II Heinz Spindler Osnabr\374cker Schriften zur Mathematik,
EHeft 3, 1994 Algebraische Geometrie I Heinz Spindler Osnabr\374cker
Schriften zur Mathematik, EHeft 5, 1995 Topologie I Rainer Vogt
Osnabr\374cker Schriften zur Mathematik, EHeft 4, 1995 \334bung
Mathematik f\374r Wirtschaftswissenschaftler, Aufgabensammlung
Roland Schw\344nzl Osnabr\374cker Schriften zur Mathematik, EHeft 2,
1993
keywords{132}:

algebra algebraische aufgabensammlung bung die einf geometrie heinz hrung mathematik nzl rainer roland schw
spindler topologie vogt

title{8}:       Scripten
url-references{326}:
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/algebra_I.ps.gz
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/algebraII.ps.gz
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/AlgGeo_I.ps.gz
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/topologie.ps.gz
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/wiwiueb.ps.gz
}
```

This content summary was generated by the Harvest system.

[Fig.2]

(Eine Verbesserung wird in 6.1 besprochen.)

Die Benutzung eines HTML-Editors ist zu empfehlen, um vorhandene HTML-Files, bzw. neu zu erstellende HTML-Files auf DTD Konformität zu überprüfen (die Harvest Distribution enthält mehrere Varianten der DTD) da syntaktisch inkorrekte Teile eines Files abgeschnitten werden. (Das Ergebnis ist im letzteren Fall eine unerwartet niedrige Ausbeute bei Brokeranfragen.)

.ps Files:
Aus PostScript-Dokumenten wird mit dem public-domain Programm ps2txt-2.1 ASCII-Code erzeugt.

Aus der Datei *ftp://esther.mathematik.uni-osnabrueck.de/pub/fizmath/fizmath.ps.gz* wird: (wir haben die folgenden Dateien stark verkürzt!)

SOIF Object for: ftp://esther.mathematik.uni-osnabrueck.de/pub/fizmath/fizmath.ps

---

@FILE { ftp://esther.mathematik.uni-osnabrueck.de/pub/fizmath/fizmath.ps
update-time{9}: 812298045
description{33}: Datenbankkommunikation unter UNIX
keywords{5670}:

abarbeitung aber agern ahigkeiten ahlt ahnlicher ahnliches ahnten ahrend aig aktionen all alle anzl datentr datum deckt del delete din dmv dort enter entfernten enth enthalten entsprechen erf erfahrungen erfolgreich ergebnissen erh erlaubt erst erw erwartet gef gefolgt gegenseitig gegenw geh gehen gelegen gesamten gesch gesellschaft indem informatik information inhaber insgesamt installation judith kann karlsruhe keine kern leistet letter letzte loginid lokalen mail man manuals math nach nachfolgende nachfolgenden nachgebildet obigen oder oglichkeit oglicht online onnen open prozess prozesse prozessen pub rechnerumgebung referate refused regelm regeln ruft rwxr scan schaffen scheinbar schnittstellenproblem schw sein seit stations statistik statistischen steht stelle stn stuttgart uberhaupt uberlassen ubermittelt ubermittlung ubrigen ucb uchtigen uck ucker uckzugreifen ugen ugt ugung uhjahr uhrende uhrt uhrtes uhrzeit ullte umer umf und userid userrechte users usr utzt variable variablen veranlat verbesserte verbindung verbindungsaufbau verborgen verbunden verbundener vereinigung verf vorangegangenen voraussetzung voraussetzungen vorg vorgehaltene vorliegenden vorliegt vorschlag vorstehenden vortrags war waren wesentlich wesentliche while wie wieder willkommensmeldung wir wird wissenschaftlich wissenschaftliche woraus worden workshop workstations wurde wurden xnr ynr zeile zeit zentralblatt zentralblatts zentrum zielgruppe zielsetzungen zugleich zugriff zum zun zur zusammen zusammenfassung zusammenhang zuse zusehen zuvor zwei zweite

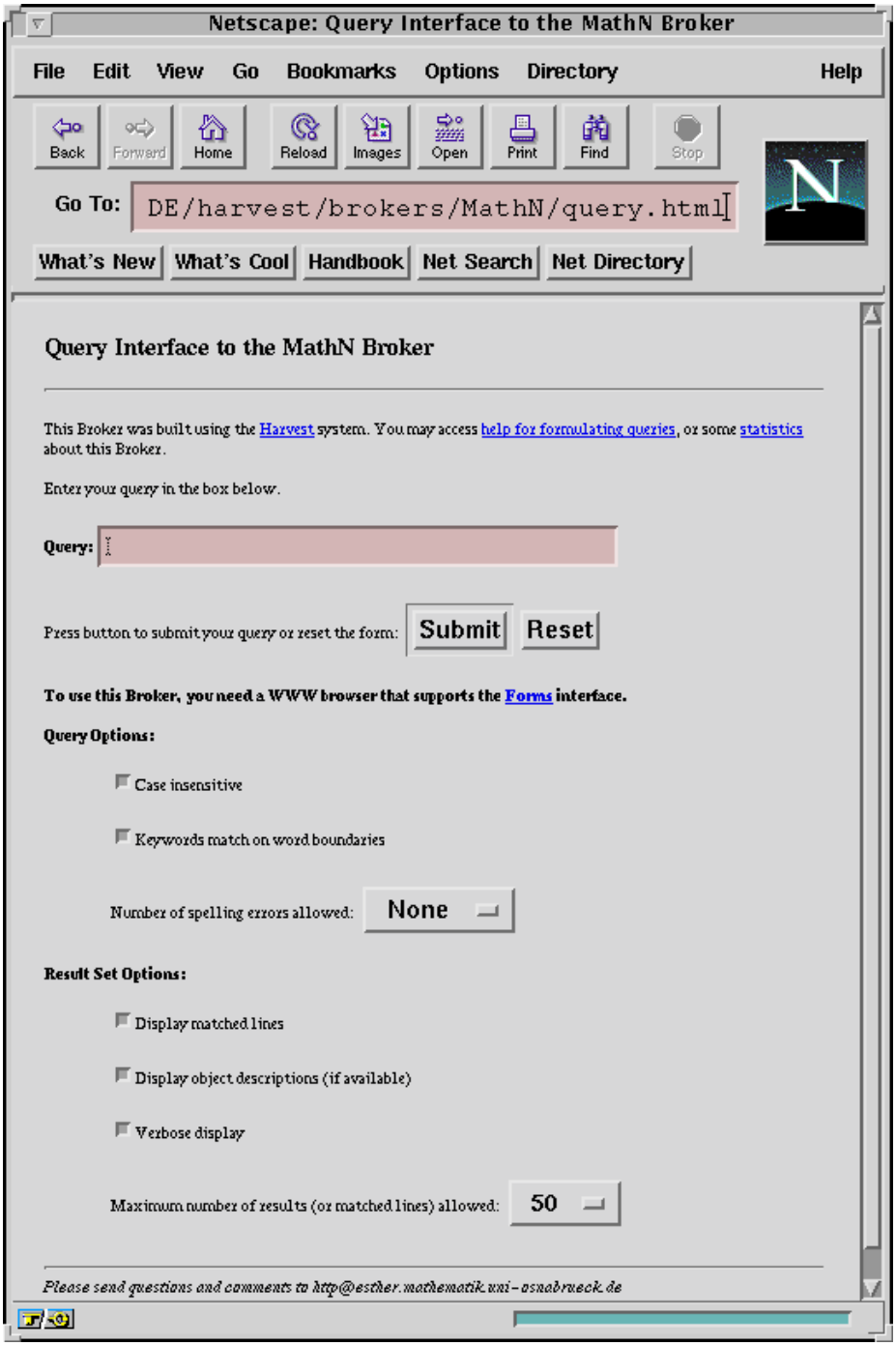
[Fig.3]

Wie man sieht, verraten die verwerteten public-domain Tools sehr deutlich ihre Herkunft aus dem **7-bit ASCII** Bereich, und es sind Anpassungen notwendig (siehe 6.1). Mit der Erstellung von Dateien im SOIF-Format ist die Arbeit des Gatherers abgeschlossen. Insgesamt ist der Gatherer die von ihrer Struktur her komplexeste Teilkomponente von Harvest. Updates des Gatherers, versuchen die Netzlast gering zu halten. Es ist jedoch nicht bei allen Access-Methoden technisch sicher möglich, festzustellen, ob ein File unverändert ist.

Der Broker

Der Broker ist im Vergleich zum Gatherer in seiner Funktionsweise erheblich einfacher. Er besteht wesentlich aus einem WWW-Form interface, einer Datenbank und einem Werkzeug zur Verbindungsaufnahme mit gewählten Gatherern und anderen Brokern. Die mit der Distribution gelieferte Datenbank ist Glimpse, die von einem Mitglied der IRTF (U. Manber) mit entwickelt wurde. Das Verbindungtool füttert bei Aufruf die vom Gatherer im SOIF-Format erstellten Dateien in die (Glimpse-)Datenbank ein. Die SOIF-Attribute

sind die Suchfelder der Anfrage auf dem Forminterface. Wird kein Suchfeld vergeben, so werden alle abgesucht. Nutzerseitig ist jederzeit *Case sensitive / Case insensitive* wählbar. Für deutschsprachige Texte ist die Wahlmöglichkeit *match on word boundaries* unüblich. Damit können Teilwörter am Anfang und Ende eines Suchstringes gefunden werden. Zusätzlich werden reguläre Ausdrücke unterstützt, was das Auffinden von Teilwörtern weiter unterstützt. Wie in Harvest nicht weiter verwunderlich, ist das Query-Interface konfigurierbar.



[Fig.4]

Da die Sources zugänglich sind, ergibt sich auch die Möglichkeit, mit etwas Programmierarbeit stärkere Menuführung des Nutzers zu erreichen. In der Ergebnismenge sind die gesuchten URL's vollständig erhalten und unmittelbar funktionsfähig.

Auf Probleme des Interfaces gehen wir unter 5.5 ein.

Wichtig bei der Konfiguration des Brokers ist die freie Wählbarkeit der zur Erstellung des Datenbankinhalts benutzten Gatherer und weiteren Broker. Es ist dadurch z. B. möglich, verschiedene Gatherer mit Spezialaufgaben zu betrauen. Da das Einsammeln von Information von Internetservern am ehesten durch Netzprobleme gefährdet ist, kann durch Verteilung der Aufgaben eine erhebliche Reduktion erforderlicher Gathererläufe erzielt werden.

Zusätzlich ist zu beachten, daß die verschiedenen Gatherer parallel und unabhängig vom Broker arbeiten können, was die Gesamt-Update-Zeit außerordentlich reduziert. Das letztliche Erstellen des Datenbankinhalts des Brokers kann (nach Grundinstallation natürlich) inkrementell erfolgen (automatisch alle 24 Stunden). Die beteiligten Gatherer stellen dem Broker die Information über Änderungen zur Verfügung.

Der Cache

Er ist wie üblich für häufig verlangte Volldokumente von Nutzen und hat auch Bedeutung für eine schnellere Bearbeitung von http-Requests durch die lokale Software. Er ist optional zu installieren und gehört nicht mehr zur Standarddistribution von Harvest. Es wäre in einem Langzeittest interessant festzustellen, ob die landläufige Hochschätzung von Object-Caches zutrifft.

Der Replicator

Der Replicator eines Brokers ist im Minimum ein mirror mit frei wählbarer Update-Zeit. Da Broker wieder von anderen Brokern eingebunden werden können, kann das Gathern über Netz vollständig unterbleiben. Die Synchronisation der Broker erfolgt dann durch Mirroring der lokal gewonnen Daten. Man beachte, daß die Topologie des Update-Graphen gewählt werden kann und damit im strengen Sinn der Netzfluß unter Nebenbedingungen optimiert werden kann. Ein erster Versuch zur Mathematisierung ist in [1] getan.

5.4 Installation von Harvest

Zur Installation eines Gatherers und eines Brokers sind keine besonderen User-Rechte erforderlich. Allerdings muß dem http-Server, unter dem der Broker laufen soll, einmalig gesagt werden, in welchem Verzeichnis die cgi-Skripten des Brokers liegen. Es werden keine Superuser-Rechte benötigt.

Man erhält sowohl den Source-Code der Software [7] als auch fertigen Binärcode für DEC's OSF/1 2.0 und 3.0, Sun OS 4.1.x und Solaris 2.3. Vom Harvest-Team nicht unterstützte Binärcodes sind erhältlich für AIX 3.2, FreeBSD, HP-UX 09.03, Linux 1.1.59 und IRIX 5.3.

Standardmäßig wird Harvest im Verzeichnis /usr/local/bin/Harvest installiert. Will man ohne Superuser-Rechte auskommen, so kann man Harvest z.B. in einem Subdirectory des http-Servers installieren. Um das Verzeichnis zu ändern, muß genau eine Zeile in einer Datei geändert werden. Genaue Angaben dazu finden sich im Manual [2]. Das Auspacken und Installieren übernimmt dann zunächst der Rechner. Diese Prozedur dauert je nach Hardwareausstattung 15 - 30 min. Falls man nicht den Standardpfad für Harvest und Perl benutzt, müssen nun noch 3 Dateien editiert werden und eine Konfigurationsdatei des zugehörigen http-Servers. Die genaue Anleitung dazu findet sich ebenfalls im Manual [2].

Die Installation zusätzlicher Komponenten ist möglich. (Etwa "Rainbow" zur Bearbeitung von .rt und .mif Dateien.)

5.5 Verhalten von Komponenten

Konfigurierbarkeit des Gatherers

Dem Gatherer wird in einem Konfigurationsfile gesagt, welche Server er besuchen soll. Er kann mit ftp-, gopher-, http-, News-, Hyper-G-Servern Kontakt aufnehmen und auch lokale Files bearbeiten. Der Administrator des Gatherers hat nun sehr viele Möglichkeiten, Server gathern zu lassen. Zunächst wird unterschieden zwischen *Root nodes* und *Leaf nodes*. *Leaf nodes* sind URL's von files, die vom Gatherer indexiert werden. Die *Root nodes* können sowohl Dokumente auf einem http-Server als auch Pfade auf Servern sein. Die *Root node* auf einen ftp-Server darf kein Dokument sein. Die Pfade beziehungsweise Links im Dokument werden vom Gatherer verfolgt. Eine *Root node* wird also vom Gatherer in mehrere *Leaf nodes* aufgespalten. Um dieses Entwickeln einer *Root node* zu kontrollieren gibt es verschiedene Möglichkeiten:

- Einstellung, wieviele *Leaf nodes* aus einer *Root node* entwickelt werden dürfen (default 250)
- Filter, die bestimmte Filenamen und Direktorynamen erlauben bzw. verbieten, auch kombinierbar in regulären Ausdrücken z.B. Tar-Archive oder README's.
- Einstellung, wieviele verschiedene Hosts besucht werden dürfen (default ist 1)
- Filter, welche Hosts besucht werden dürfen (z.B. nur Rechner mit Internetnummer 131.173.*)
- die Tiefe, wie weit die Verzweigung einer *Root node* sein darf (default unendlich)
- ob auch Links auf andere Servertypen verfolgt werden sollen (z.B. von http- auf einen ftp-Server)

All diese Einstellungen sind im Manual [2] gut beschrieben und funktionieren wie dargestellt.

Summarizer

Die Summarizer sind zum Teil public-domain Tools, vom Harvestteam programmiert oder auch kommerzielle Software (Rainbow). Ein Problem dabei ist der Umgang mit Umlauten. Sie werden zum Teil nicht dargestellt und die Wörter werden auseinandergehackt (.ps), oder sie werden ohne Konfigurationsarbeit als ASCII-Nummer dargestellt (.html). Das Zerhacken von Wörtern findet sich im Beispiel [Fig.3]. Der .ps-Summarizer besteht aus einem Shell-Script und C-Programmen (siehe 6.1).

Broker

Das Brokerinterface kommuniziert über Formwidgets mit der Datenbank, die vom Broker aus den Gather-Daten gemacht wird. Das erste Problem dabei sind wieder die Umlaute. Die einzige Möglichkeit nach einem Wort mit einem Umlaut zu suchen, ist, den Umlaut per Tastendruck einzugeben. Dieses ist auf einer US-Tastatur nur nach Umbelegen von einzelnen Tasten möglich. Bei Eingabe von Umlauten in .html Syntax bekommt man nur eine Fehlermeldung.

Weiter gibt es im Brokerinterface die Taste *Keywords match on word boundaries*, die genau umgekehrt zur Erwartung des Nutzers arbeitet. Ebenfalls verwunderlich ist, daß die Einstellung *bis zu 2 Fehlern* bei einem Query eine kleinere Menge an Artikeln matched als die Einstellung *0 Fehler*.

Diese Fehler liegen in cgi-Skripten und sind daher vermutlich auch durch Nutzer zu beheben.

6 Der MathN Broker

Der MathN Broker *http://www.mathematik.uni-osnabrueck.de/harvest/brokers/MathN/* indexiert die Preprints und Skripten der mathematischen Serverstandorte (insgesamt 38 Einzelserver), die auf der nachfolgend wiedergegebenen "Sensitive Map" eingetragen sind.

[Fig.5]

Dazu sind lokal 5 Gatherer installiert. Neben dem Broker des Osnabrücker Servers sind Gatherer aus Kaiserslautern und Rostock eingebunden.

Der MathN Broker weist derzeit ca. 2800 Dokumente nach. Durch Patchen des PostScript- und des html-Summarizers können die Anfragen in vernünftiger Form gestellt werden. Die Antwortzeiten des Brokers sind für einfache Anfragen akzeptabel. Es handelt sich derzeit um eine Installation von Harvest-1.4.pl1 unter SunOS 4.1.3 / Perl 4. Da er nur auf einer Sparc 1+ / 32 MB Hauptspeicher arbeitet, wird das Abfrageverhalten bei boolschen Anfragen schnell schlechter (Wartezeiten von einer halben Minute geben noch keinen Anlaß am Erfolg der Anfrage zu zweifeln). Die Variablen: Prozessorleistung und Hauptspeicherausstattung schlagen gnadenlos zu. Der Broker wird in Kürze auf einer leistungsfähigeren Maschine arbeiten.

Für das Einsammeln der Information ist die Netzkapazität von ausschlaggebender Bedeutung. Es ist für den Nutzer weitgehend uninteressant, ob die Aufbereitung der Daten 4, 8 oder 12 Stunden in Anspruch nimmt. Die CPU-Leistung ist wieder wesentlich für den Moment des (inkrementellen) Updates des Brokers relativ zu den gesammelten Daten, da während der Update-Zeit der Broker unzugänglich ist.

Bei Einsatz der Replicatorentechnik könnte die Netzbelastung drastisch reduziert werden: Gathern sollte vollständig lokal erfolgen. Ein zentraler Broker - oder nur seine Konfiguration - sollten lokal gespiegelt werden. Eine Konfiguration dieser Art optimiert die Netzlast, minimiert Probleme bei lokalen Netzausfällen, erlaubt Kontrolle über die Aktualität der Daten und liefert vernünftige Antwortzeiten auch bei komplexen Queries.

6.1 Umlaute

Das Fehlverhalten bzgl. Umlauten ist auf MathN derzeit folgendermaßen behoben: Der .ps Summarizer wurde gepatcht. Für die Wirkung dieser schlichten Modifikation des zugehörigen Programmes (harvest-1.4.pl1/components/gatherer/standard/ps2txt/ps2txt-2.1.c) siehe folgendes Objekt des Osnabrücker Gatherers: (wir haben uns entschieden Umlaute auf Grundvokal mit angehängtem “e” abzubilden, Darstellung als Umlaute ist auch möglich):

SOIF Object for:
ftp://esther.mathematik.uni-osnabrueck.de/pub/fizmath/fizmath.ps.gz

---

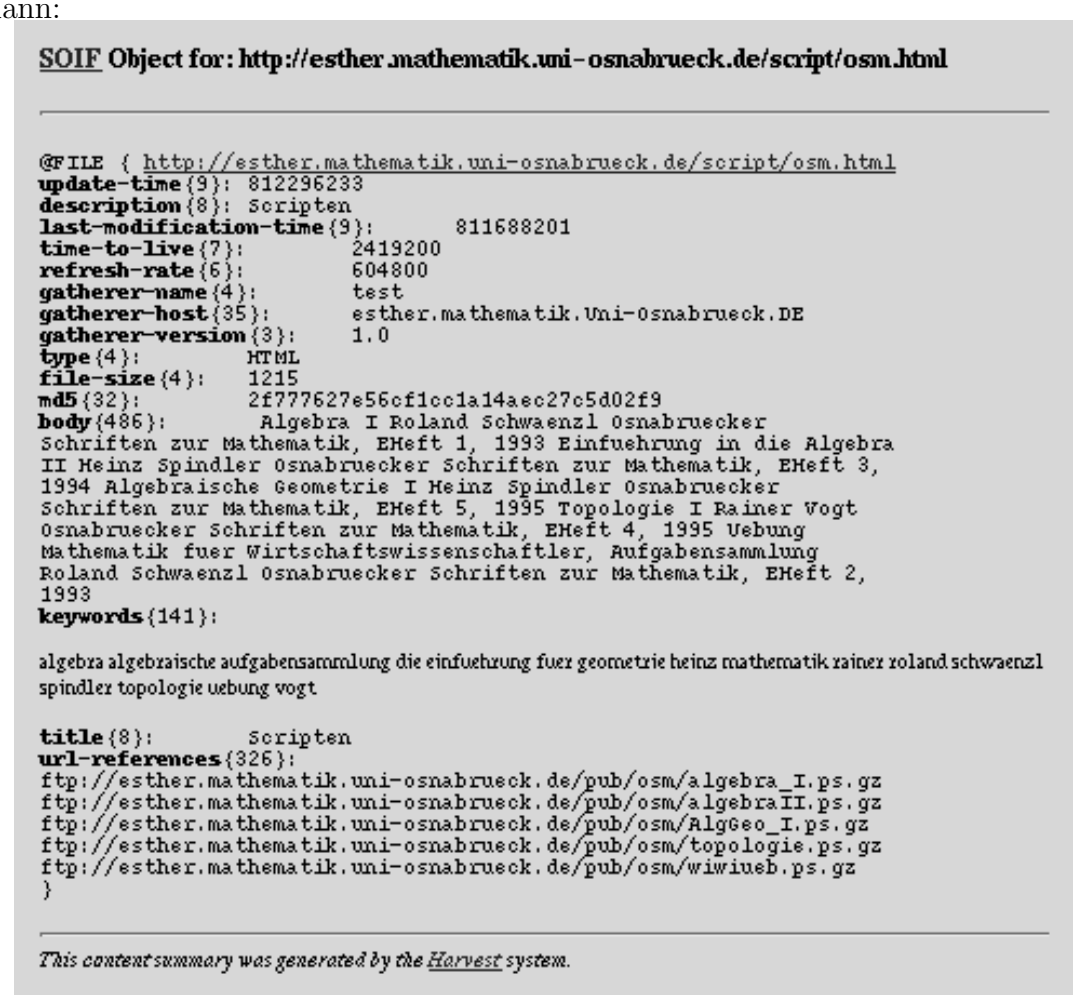
@FILE { ftp://esther.mathematik.uni-osnabrueck.de/pub/fizmath/fizmath.ps.gz
update-time{9}: 812298015
embed<1>-type{10}: PostScript
embed<1>-keywords{7055}:

abarbeitung aber abgefangene abgesetzt ablauf accounting adresse aktionen all antwort anwendung arbeiten arbeitsplaetze daraus das dass dat date dateaus datei dateienthaelt durch durchgefuehrtes durchlaufen ebenso eigene eigentlichen ein eine einem einen einer eines einfacher eingabe eingebaute einheitliche entsprechen entsprechenden entstanden entwickelt entwickelte erfahrungen erfolgreich erfolgreichen erneute erreicht erscheinen ersetzen erst explizite fachbereiche fachinformation fachinformationszentrums fest file fileein filtern fruehjahr ftp fuehrt fuer funktionstuechtigen gefoerdert gegenwaertige gehoert genuegt geschaeftsstelle geschuetzt giltaehnliches groessten haette informatik information inhaber insgesamt jahre lichen listing local loesung log moeglichkeit nach naechstes naemlich osnabrueck osnabruecker out platz pluemer porten return roland rom ruft schwaenzl script seit selbst simuliert sind sinnvoll sitzung skizziert skript skripts software sollen tclwurden tcp tech technische technischen technologie teil teilen teilmenge telnet telnetdie telnetnach terminal tests textsprache trying turin ucb ueber ueberarbeitete uhrzeit umarbeitung umfaengliches umgebenden und undaehnlicher uni unirech unirechkonnte unix worden work workshop writing wuenschen wuerde wurde zedur zeduren zeile zeit zentralblatt zentralblatts zentrum zielgruppe zielsetzungen zugang zugleich zugriff zum zunaechst zunaechstueberhaupt zur zurueckzugreifen zusammen zusammenfassung zuse zusehen zuvor zwei zweite zweiten

[Fig.6]

(Patch von ps2txt: [8])

Der .html- bzw. sgml-Summarizer kann auf Sun-Betriebssystemen durch setzen von LC_CTYPE=iso_8859_1 in der Datei "RunGatherer" des jeweiligen Gatherers Umlaute direkt darstellen. Wir haben sie derzeit auf Grundvokal mit angehängtem "e" abgebildet, dazu ist eine Änderung von SGML.sum ausreichend [8]. Aus dem Dokument [Fig.1] wird dann:



SOIF Object for: http://esther.mathematik.uni-osnabrueck.de/script/osm.html

---

```
@FILE { http://esther.mathematik.uni-osnabrueck.de/script/osm.html
update-time{9}: 812296233
description{8}: Scripten
last-modification-time{9}:      811688201
time-to-live{7}:        2419200
refresh-rate{6}:        604800
gatherer-name{4}:       test
gatherer-host{35}:      esther.mathematik.Uni-Osnabrueck.DE
gatherer-version{3}:    1.0
type{4}:        HTML
file-size{4}:   1215
md5{32}:        2f777627e56cf1cc1a14aec27c5d02f9
body{486}:       Algebra I Roland Schwaenzl Osnabruecker
Schriften zur Mathematik, EHeft 1, 1993 Einfuehrung in die Algebra
II Heinz Spindler Osnabruecker Schriften zur Mathematik, EHeft 3,
1994 Algebraische Geometrie I Heinz Spindler Osnabruecker
Schriften zur Mathematik, EHeft 5, 1995 Topologie I Rainer Vogt
Osnabruecker Schriften zur Mathematik, EHeft 4, 1995 Uebung
Mathematik fuer Wirtschaftswissenschaftler, Aufgabensammlung
Roland Schwaenzl Osnabruecker Schriften zur Mathematik, EHeft 2,
1993
keywords{141}:
```

algebra algebraische aufgabensammlung die einfuehrung fuer geometrie heinz mathematik rainer roland schwaenzl spindler topologie uebung vogt

```
title{8}:       Scripten
url-references{326}:
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/algebra_I.ps.gz
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/algebraII.ps.gz
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/AlgGeo_I.ps.gz
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/topologie.ps.gz
ftp://esther.mathematik.uni-osnabrueck.de/pub/osm/wiwiueb.ps.gz
}
```

This content summary was generated by the Harvest system.

[Fig.7]

6.2 Administration

Der laufende Betrieb des MathN Brokers verursacht keinen nennenswerten Zeitaufwand. Der Broker wird von cron automatisch wieder gestartet, falls der Server gebootet wird. Die Gatherer werden jeweils zweimal im Monat zeitversetzt von cron gestartet und geben die aktualisierten Daten automatisch an den Broker weiter.

Arbeit für den Administrator entsteht, wenn die Liste der zu gathernden Server erweitert werden soll. Der Zeitaufwand dafür hängt erheblich davon ab, wie der zu gathernde Server organisiert ist.

Wenn nur ein Directory eingesammelt werden muß, ist die Konfiguration schnell und sauber zu erledigen:

```
http://www.zib-berlin.de/ZIBbib/ URL=450,ZIB-Filter
```

Die Filterdatei hat folgenden Inhalt:

```
Allow .*/ZIBbib/.*
Deny .*
```

Liegen die Dokumente vor Ort auf verschiedenen Servern, sollte es ein Index-File geben, von dem aus alle Dokumente erreicht werden können. Wünschenswert ist es, daß die Dokumente jeweils in einem Subdirectory mit einem einheitlichen Namen liegen, z.B. `/papers/`. Durch eine Zeile in einem Filter für den Gatherer kann dann verhindert werden, daß z.B. ganze Homepages von Fachbereichen indexiert werden.

Ein weniger gutes Beispiel ist Uni X:

```
http://care.mathematik.uni-X.de/mozilla/publications.g.html DEPTH=1
ftp://less.uni-X.de/pub/
http://angel.mathematik.uni-X.de/papers/ URL=250,X-Filter
```

An diesem Standort gibt es fünf verschiedene Server. Um die Konfiguration des Gatherers zu schreiben, müssen alle diese fünf Arbeitsgruppen-Server besucht werden, um festzustellen, daß nur drei von ihnen überhaupt Preprints anbieten. Die Dokumente sind im zweiten Fall auf einem ftp-Server, im dritten Fall in Unterverzeichnissen mit Abstracts in TeX-Form und im ersten Fall nach Jahrgängen geordnet in Unterverzeichnissen mit dem Vornamen des jeweiligen Autors als .dvi-Files oder in PostScript gespeichert.

Sicherlich hat jede dieser gewählten Darstellungen ihre Vorteile. In der vorliegenden Form verursachen sie durch Häufung vermeidbaren Zeitaufwand. Auf einen Ausweg unter Wahrung der individuellen Gestaltungswünsche werden wir unter 8. eingehen.

7 Einsetzbarkeit von Harvest

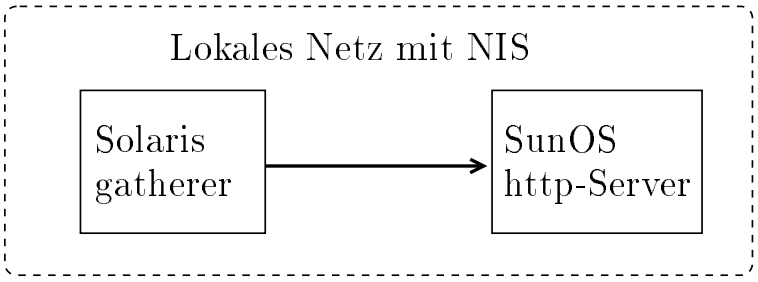
Die in Abschnitt 6. beschriebene Konfiguration läuft stabil unter Harvest 1.4pl1 auf SunOS 4.1.3 mit Perl 4 als Skriptinterpreter. Dabei befindet sich der http-Server auf derselben Maschine (Sparc 1+, 32 MB RAM). Erfolgreich verlief unter Solaris 2.4 die Indexierung einer großen, entfernt liegenden Site, die über DNS kontaktiert wurde (Harvestinstallation auf: Sparc 20, 64 MB RAM).

Über das lokale Gathern findet man eine Bemerkung im User Manual (p. 24):

"Some sites may use Sun Microsystem's Network Information Service (NIS) instead of, or in addition to, DNS. We believe that Harvest works on systems where NIS has been properly configured. The NIS servers (the names of which you can determine from the `ypwhich` *command) must be configured to query DNS servers for hostnames they do not know about. See the -b option of the* `ypxfr` *command.*

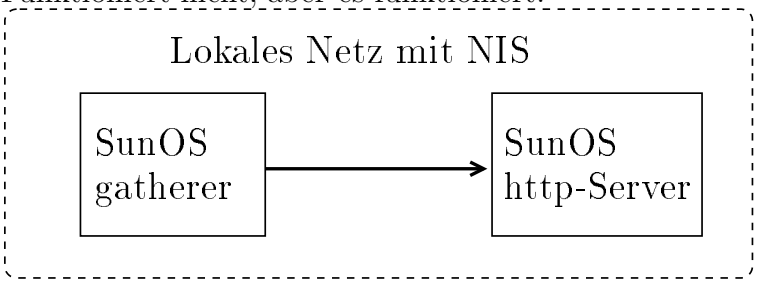
We would welcome reports of Harvest successfully working with NIS. Please email us at harvest-dvl@cs.colorado.edu."

Diese dunkle Bemerkung läßt sich dahingehend präzisieren:



[Fig.8]

Funktioniert nicht, aber es funktioniert:



[Fig.9]

Das Gatherern von WWW-Servern außerhalb des lokalen Netzes ist in beiden Fällen unproblematisch. Für ein optimales Laufzeitverhalten ist es in jedem Fall sinnvoll, Gatherer und http-Server auf derselben Maschine zu installieren.

Wie aus der Newsgroup `comp.infosystems.harvest` (archiviert in [4]) zu entnehmen ist, wird Harvest auch unter HP-, SGI-, IBM- und Dec-Unixen eingesetzt. Installationen unter Linux werden ebenfalls erwähnt.

Dort wurde ausgeführt, daß ein spanning tree Algorithmus zum Auflisten der Sites auf http-Servern bei Aktivierung der Option `Depth` falsch wird. Es werden dann systematisch Files übersehen. Ein Patch ist angekündigt.

Die Handhabung des "Umlaut" Problems wurde schon im vorigen Paragraphen besprochen.

Überarbeitet werden muß das Query-Interface des Brokers. Die Bezeichnungen der Schalter für die Optionen sind verwirrend.

Derzeit wenig sinnvoll ist die Einführung von Suchfeldern wie Author oder MSC-Klassifikation, da die entsprechenden Items in aller Regel aus den Rohdaten nicht sicher extrahiert werden können.

Boolsche Suche nach Stichwörtern ist jedoch heute schon ein nützliches Instrument.

Bei großen Dokumentzahlen ($> \mathbf{65000}$) muß eine mächtigere Datenbank, als die in der Lieferung enthaltene public domain Datenbank "glimpse" eingebaut werden. Das Manual gibt Hinweise, daß dies möglich ist. Der Einsatz von WAIS-Versionen wird besprochen.

Bei der Ausdehnung der Indexierungbasis auf ein größeres zeitliches und/oder geographisches Gebiet wird der Einbau einer leistungsfähigeren Datenbank vordringlich sein. Als limitierender Faktor darf wie stets auch die schnell wachsende Rechnerbelastung bei komplexer boolscher Suche nicht außer Acht gelassen werden.

Die Gruppe der Entwickler an der University of Colorado hat sich in den letzten Monaten erheblich verkleinert. Die Softwareevolution wird sich folglich verlangsamen.

In diesem Zusammenhang ist das bei der Gesellschaft für Informatik vorhandene generelle Interesse an der (Weiter-)Entwicklung von Brokern für die Strukturierung mathematischer Daten relevant.

8 Verbesserung der Dokumentationsmöglichkeiten

Wie schon unter 6. erwähnt, gibt es an den Fachbereichen der Bundesrepublik divergierende Vorstellungen über die Darstellung von Dokumenten im WWW.

Das Eigenformat zur Speicherung von Texten auf einem WWW Server ist natürlich HTML. Die derzeitig verbreitete Version HTML2 ist jedoch für mathematische Texte unzureichend.

Häufig werden TeX- und dvi-Files angeboten. Dabei kann es aber zu Problemen bei der Übersetzung oder beim Previewing kommen: Nutzung unterschiedlicher Style-Files, Macro packages und Fonts, nicht zu vergessen die verschiedenen TeX-Dialekte.

Grundsätzlich sollten daher mathematische Texte in PostScript Format - zur Reduktion der Übertragungszeiten in gezippter Form - angeboten werden.

Zur Indexierung von Preprints, Skripten und Software ist es vorteilhaft, wenn es

- ein Dokument gibt, von dem aus alle Preprints und Skripten erreicht werden können. Wenn
- die Dokumente jeweils in Subdirectories liegen, die einen einheitlichen Namen - z.B. `/papers/` - haben.

Die Schnittmenge aller Dokumente eines Fachbereiches, die diese genannten Punkte erfüllen, sollte genau die Menge an Dokumenten ergeben, die vom Preprint-Gatherer erfaßt werden sollen. Dabei muß das Dokument, von dem aus man alle mathematischen Dokumente erreicht, nicht notwendigerweise von der Homepage des Fachbereichs aus zugänglich sein, es dient nur dem Gatherer.

Durch die genannten Punkte wird daher die Gestaltungfreiheit der Serverbetreiber nicht beschränkt. Es werden allenfalls einige Directories umbenannt und ein neues Dokument hinzugefügt.

Der Administrationsaufwand wie unter 6.2 beschrieben, wird auf diese Weise minimiert.

Zur Verbesserung der Indexierungsergebnisse kann der META-tag unter HTML benutzt werden. Der META-tag ermöglicht es, einem html-File **unsichtbare** Informationen mitzugeben. Das HTML-File:

```
<HEAD>
<TITLE>Meine kleine Welt</TITLE$>
</HEAD>
<BODY>
Hallo
</BODY>
```

Erscheint unter einem WWW-Browser genauso wie das File:

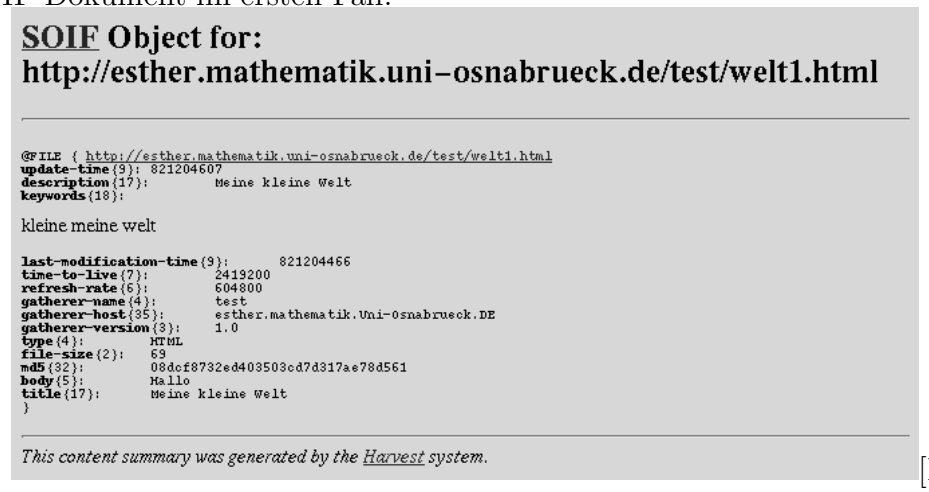
```
<HEAD>
<TITLE>Meine kleine Welt</TITLE>
<META NAME=‘‘Author’’ CONTENT=‘‘Erika Musterfrau’’>
<META NAME=‘‘Author’’ CONTENT=‘‘Karl Mustermann’’>
</HEAD>
<BODY>
Hallo
</BODY>
```

Im zweiten Fall kann der Gatherer in seiner Datenbasis die Autoren spezifizieren, was im ersten Fall natürlich nicht möglich ist.

Bild im Browser:

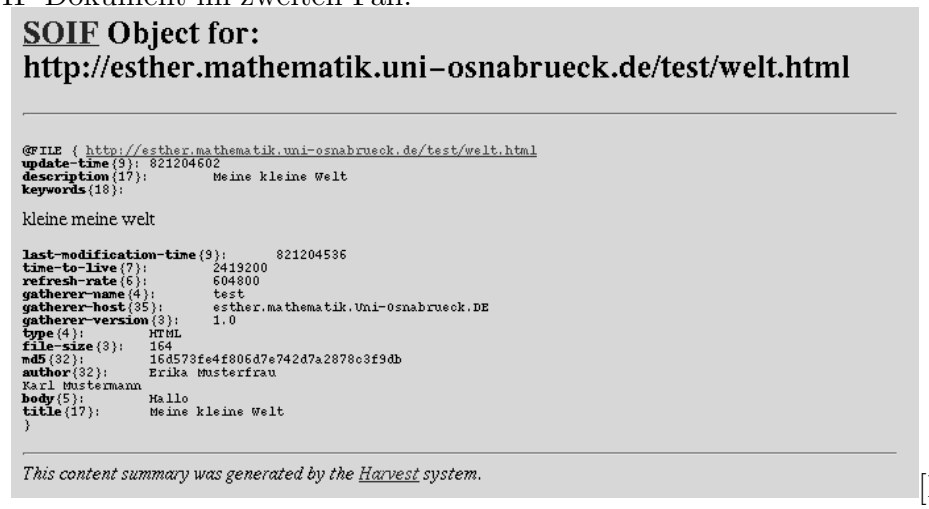
[Fig.10]

SOIF Dokument im ersten Fall:



[Fig.11]

SOIF Dokument im zweiten Fall:



[Fig.12]

Da aus PostScript-Files Informationen wie Autor und Titel nicht sicher automatisch extrahiert werden können, sollte der http-Administrator zu jedem PostScript-Dokument ein knappes html-File mit META-tags verfassen. Diese Begleitfiles könnten beispielsweise in einem Unterverzeichnis von `/papers/` mit dem Namen `/papers/shadows/` gesammelt abgelegt werden. Konkret könnte das wie folgt aussehen:

Zum gepackten File algebra1.ps.gz des Beispiels [Fig.1] assoziiere man:
algebra1.shadow.html:

```
<HEAD><TITLE>Algebra I</TITLE>
<META NAME=‘‘Author’’ CONTENT=‘‘Roland Schwnzl’’>
<META NAME=‘‘MSC’’ CONTENT=‘‘16-01’’>
<META NAME=‘‘Year’’ CONTENT=‘‘1993’’>
<META NAME=‘‘Update’’ CONTENT=‘‘Nov 28 1995’’>
<META NAME=‘‘Series’’ CONTENT=‘‘Osnabrücker Schriften zur Mathematik’’>
<META NAME=‘‘Subseries’’ CONTENT=‘‘Vorlesungsskripten’’>
</HEAD>
<DL>
<DT><A HREF=‘‘ftp://ftp.mathematik.uni-osnabrueck.de/pub/osm/algebra_I.ps.gz’’>
<STRONG> Algebra I </STRONG></A>
<DD><STRONG>Roland Schwnzl</STRONG>
<DD>Osnabrcker Schriften zur Mathematik, EHeft 1, 1993
</DL>
</BODY>
```

Als SOIF Dokument entsteht daraus:

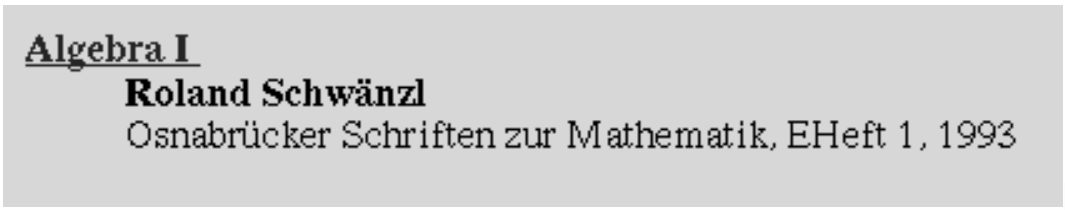
SOIF Object for:
http://esther.mathematik.uni-osnabrueck.de/test/osm.html

```
@FILE { http://esther.mathematik.uni-osnabrueck.de/test/osm.html
update-time{9}: 821197859
description{9}: Algebra I
last-modification-time{9}:      821197821
time-to-live{7}:        2419200
refresh-rate{6}:        604800
gatherer-name{4}:       test
gatherer-host{35}:      esther.mathematik.Uni-Osnabrueck.DE
gatherer-version{3}:    1.0
type{4}:        HTML
file-size{3}:   567
md5{32}:        384f81cf0df58121763da9f15ef357cb
author{16}:     Roland Schwaenzl
body{80}:        Algebra I Roland Schwaenzl Osnabruecker Schriften zur Mathematik,
EHeft 1, 1993
keywords{25}:

algebra roland schwaenzl

msc{5}: 16-01
series{37}:     Osnabruecker Schriften zur Mathematik
subseries{18}:  Vorlesungsskripten
title{9}:       Algebra I
update{11}:     Nov 28 1995
url-references{62}:     ftp://ftp.mathematik.uni-osnabrueck.de/pub/osm/algebra_I.ps.gz
year{4}:        1993
}
```

This content summary was generated by the Harvest system.

[Fig.13]

Im Browser erscheint das Dokument:

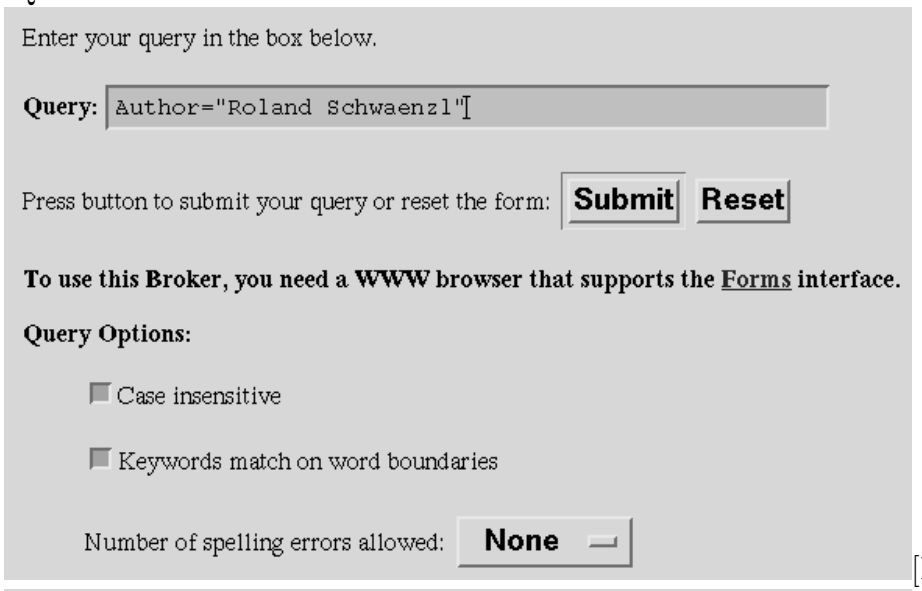


[Fig.14]

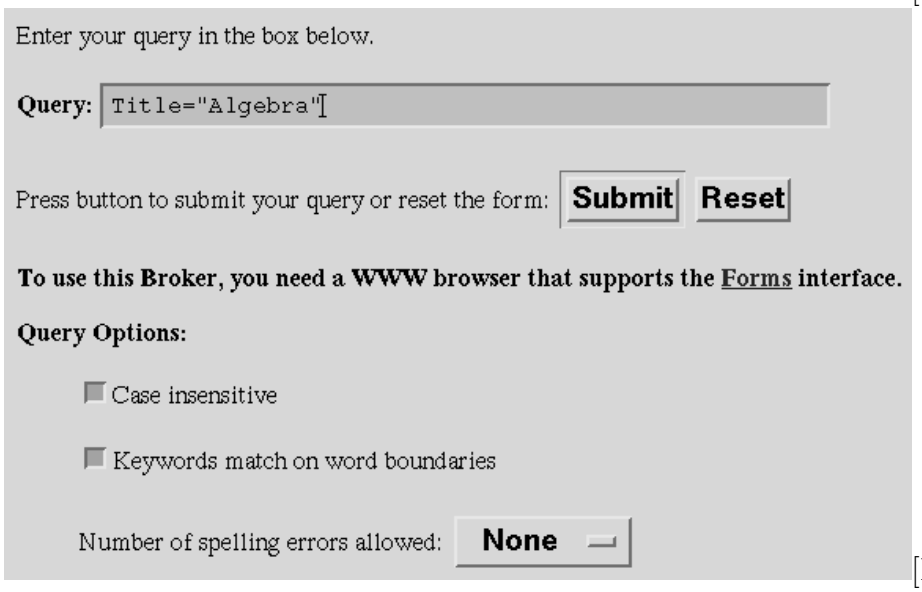
Dem Serverbetreiber hat es in diesem Beispiel gefallen, die MSC Klassifikation zwar als META-tag im SOIF Dokument festhalten zu lassen, aber dem menschlichen Auge wollte er die Nummer nicht präsentieren.

Auf dem Server muß kein **sichtbarer** Link auf Schattenfiles vorhanden sein. Das Erscheinungsbild eines Servers wird nicht beeinträchtigt.

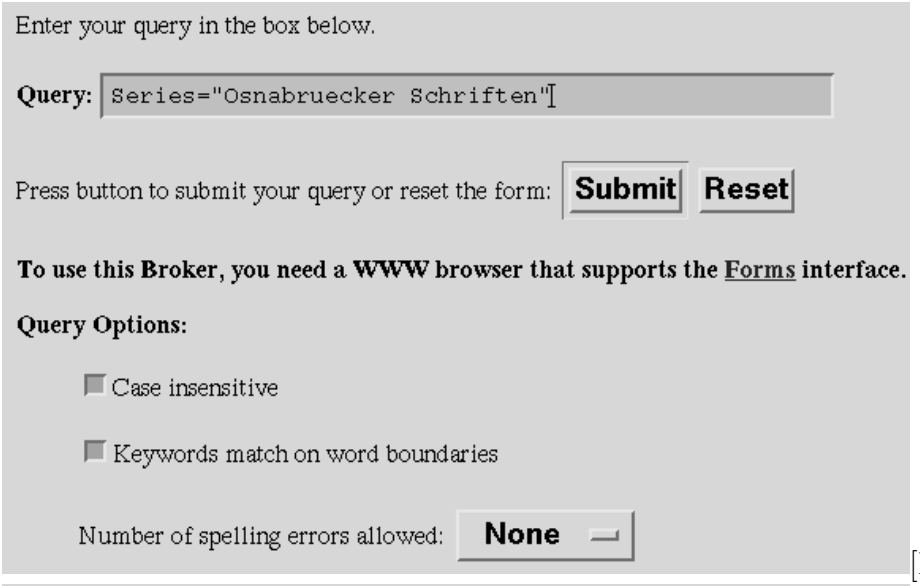
Die Wohlüberlegtheit der Konstruktion von Essence und Broker erlaubt nun das Schattenfile - und damit schließlich auch das eigentliche Dokument - mit jedem der folgenden Queries aufzufinden:



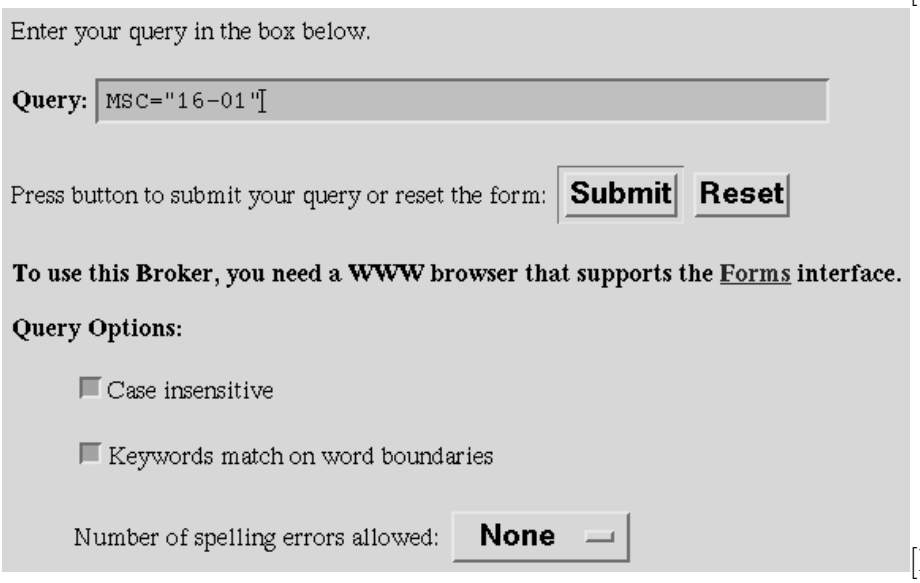
[Fig.15]



[Fig.16]



[Fig.17]



[Fig.18]

Dafür ist **keine** Modifikation der Harvest Software erforderlich. Ein wertvolles Instrument werden Schattenfiles aber nur dann, wenn sie im Bereich der zu gathernden Server allgemein und mit einem identischen Grundbestand von verbindlichen, einheitlich bezeichneten META-tags verwendet werden.

Die American Mathematical Society benutzt den folgenden Katalog [9] von Zusatzeinträgen für ihren Preprintserver.

```
% Template for submission of a Mathematical Preprint Abstract
% to the AMS Preprint Server on e-MATH     Revision date: April 27, 1995
%
% Percent sign at beginning of line indicates comments.
% Indent carryover lines with whitespace (space or tab).
%
% After filling out this form, send to:  e-prints@e-math.ams.org
```

```
%
%
Type: Preprint Submission
% Mandatory field.

Title:
% Mandatory field.

Author:
% Mandatory field.
% Last Name, First Name
% Use multiple Author: lines for each author.
% Author institutions and addresses (optional) belong in the "Notes:" below.

Pclass:
Sclass:
% Primary and Secondary Mathematics Subject Classifications.
% Primary is a mandatory field.  Secondary is an optional field.

Keywords:
%Keywords from preprint.  Optional field.

Contact:
% E-mail address for contact person.
% Mandatory field.

Expiration-Date:
% Date when preprint will be removed from AMS Preprint server. Optional field.
% Default date is two years from date of submission.

TeX-Type:
% Choices: AMS-TeX, AMS-LaTeX, plain TeX, LaTeX, LAMSTeX.
% Mandatory field if submission includes TeX file(s).

URL-Pointer:
% Standard URL syntax.
% Mandatory field if Non-URL-Pointer: and Main-TeX-Filename: are blank.

Non-URL-Pointer:
% Use only if your preprint is not available electronically.
% Mandatory field if URL-Pointer: and Main-TeX-Filename: are blank.

Copyright:
```

```
% Copyright statement. Optional field.
% Default statement is "Full copyright maintained by authors."

Begin-Abstract:

End-Abstract:
% Both Abstract fields are mandatory.  Entry may be several lines.

Notes:
% May contain author(s) address(es), publisher information, preprint series
% information, pointers to related materials. Blank line denotes end of
% notes information.  Optional field.

%-------------Cut here if submitting an abstract only.-------------

Main-TeX-Filename: filename.ext
% For submitting TeX files for full preprint texts, specify the name of
% the "main" TeX file here, meaning the file on which a TeX processor should
% be run.  Mandatory field if URL-Pointer: and Non-URL-Pointer: are blank.

Begin-File: filename.ext
% Specify the filename above and insert the file here.

End-File:
% Repeat Begin-File:   and   End-File: tags for each included file.

% For more details on submitting preprints electronically, see the WWW page at
%       http://www.ams.org/web/preprints/submission-instructions.html
```

Tag's für die Lebensdauer sind für die in unserem Kontext zu indexierenden Server entbehrlich, da Dokumente nie den Bereich unmittelbarer Zugriffsmöglichkeit des Autors verlassen. Eine abzustimmende **Auswahl** der verbleibenden Eingänge sollte jedoch auch für den hiesigen Kreis von Servern praktikabel sein.

Im Body eines Schattens könnte auch ein knappes vom Autor verfaßtes Abstract einer Arbeit abgelegt werden.

Durch diese "unsichtbaren" Änderungen in der Bestückung von WWW-Servern wird Harvest eine Datenbank mit suchbaren bibliographischen Daten erzeugen, zusätzlich zu den Eingängen, die heute bereits über summarizing verfügbar sind.

Ein Umschreiben des Query-Interfaces würde dann Harvestbroker auch für Nutzer ohne längere Eingewöhnung in die Datenbanksprache zu einem Werkzeug mit hoher Zielgenauigkeit machen.

Durch Traversieren nach MSC Klassifikation werden Profildienste denkbar, die ohne Umorganisation der Grundstrukturen auskommen.

References

1. P. Danzig, K. Obrazka, D. Delucia, and N. Alam: Massively replicating services in autonomously managed wide-area Internet works. TR, Dept. Compter Sci., Univ. Southern California.
   ftp://catarina.usc.edu/pub/kobraczk/ton.ps.z

2. D.R. Hardy, M. F. Schwartz, D. Wessels: Harvest Users's Manual, TR CU-Cs-743-94, Sept. '95, Univ. Colorado at Boulder
   ftp://ftp.ask.uni-karlsruhe.de/pub/infosystems/harvest/harvest-1.3-docs.tar.gz

3. M. Grötschel, J. Lügger: Aufbau elektronischer Informations- und Kommnikationsstrukturen Deutscher Dokumentartag, Proceedings einer Konferenz an der Fh. Potsdam, 26.-28.9.95, Deutsche Gesellschaft für Dokumentation, Frankfurt, 1995, 13-58.

4. http://harvest.cs.colorado.edu/Harvest/brokers/CIH/

5. A. Frühbis: Preprintbereitstellung und Zugriff über Harvest.
   http://www.mathematik.uni-osnabrueck.de/workshop/vortraege/fruehbis/fruehbis.html

6. Gatherer: hp710.math.uni-rostock.de:8500

7. http://harvest.cs.colorado.edu/harvest/gettingsoftware.html

8. http://www.mathematik.uni-osnabrueck.de/harvest/gatherers/berlin/bin/PostScript.sum
   http://www.mathematik.uni-osnabrueck.de/harvest/gatherers/berlin/bin/psosna.c
   http://www.mathematik.uni-osnabrueck.de/harvest/gatherers/berlin/bin/SGML.sum

9. http://www.ams.org/preprints/template.fil

Judith Plümer, judith@scarlett.mathematik.uni-osnabrueck.de
Roland Schwänzl, roland@scarlett.mathematik.uni-osnabrueck.de

Fachbereich Mathematik / Informatik
Universität Osnabrück
Albrechtstraße 28
49069 Osnabrück