Harvest: Efficient, Interoperable Indexing & Searching

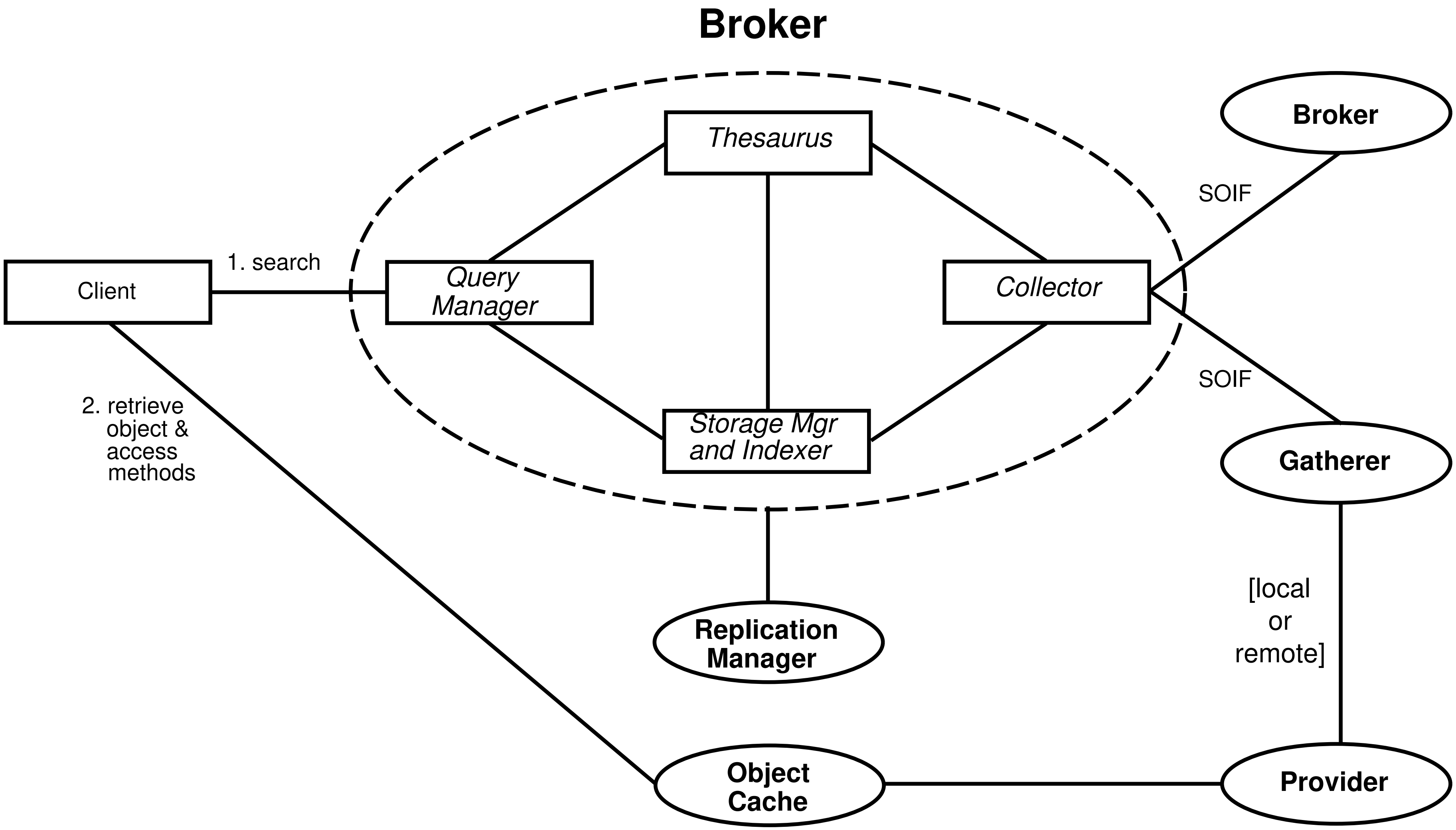
IRTF-RD:

Mic Bowman	(co-PI, Transarc)
Peter Danzig	(co-PI, USC)
Udi Manber	(co-PI, U Arizona)
Mike Schwartz	(PI, U Colorado)

PRAs:

Darren Hardy	(U Colorado)
Duane Wessels	(U Colorado)

+ many students


Broker
Thesaurus
Query Manager
Collector
Storage Mgr and Indexer
Client
1. search
2. retrieve object & access methods
Broker
SOIF
SOIF
Gatherer
Replication Manager
[local or remote]
Object Cache
Provider

Gatherer specification:

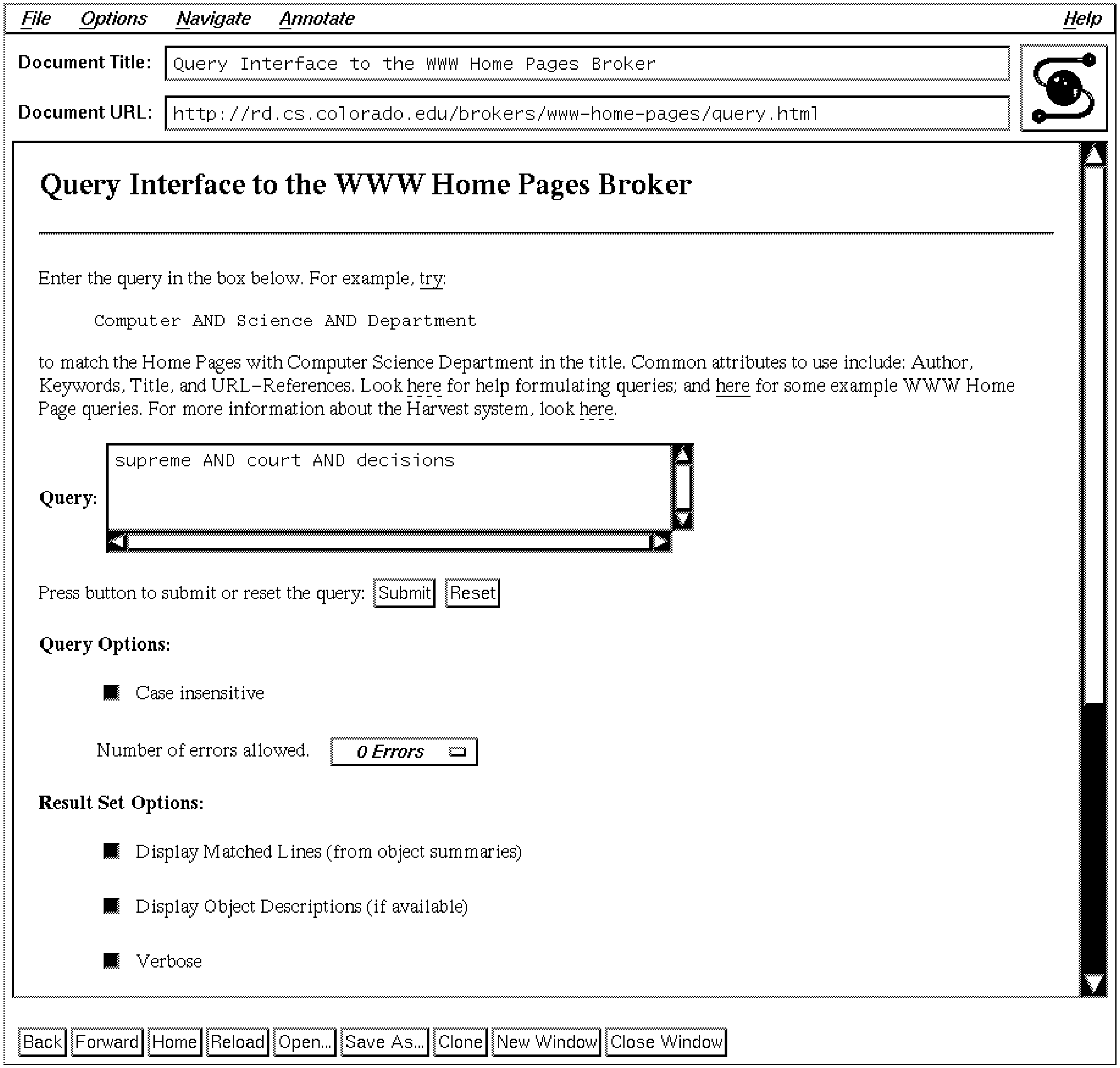
```
<RootNodes>
ftp://ftp.cs.colorado.edu/pub/cs/techreports/hardy
http://harvest.cs.colorado.edu/
</RootNodes>

<LeafNodes>
http://powell.cs.colorado.edu/~hardy/Home.html
ftp://ftp.cs.colorado.edu/pub/cs/distribs/essence/Essence.tar.Z
</LeafNodes>
```

Gatherer actions:

- enumeration
- retrieval
- unnesting
- type-specific summarizing
- serving incr.-updated, compressed summary streams

Broker actions:

- SOIF stream retrieval
- indexing
- MD5-based duplicate elimination
- serving (HTML) query interface

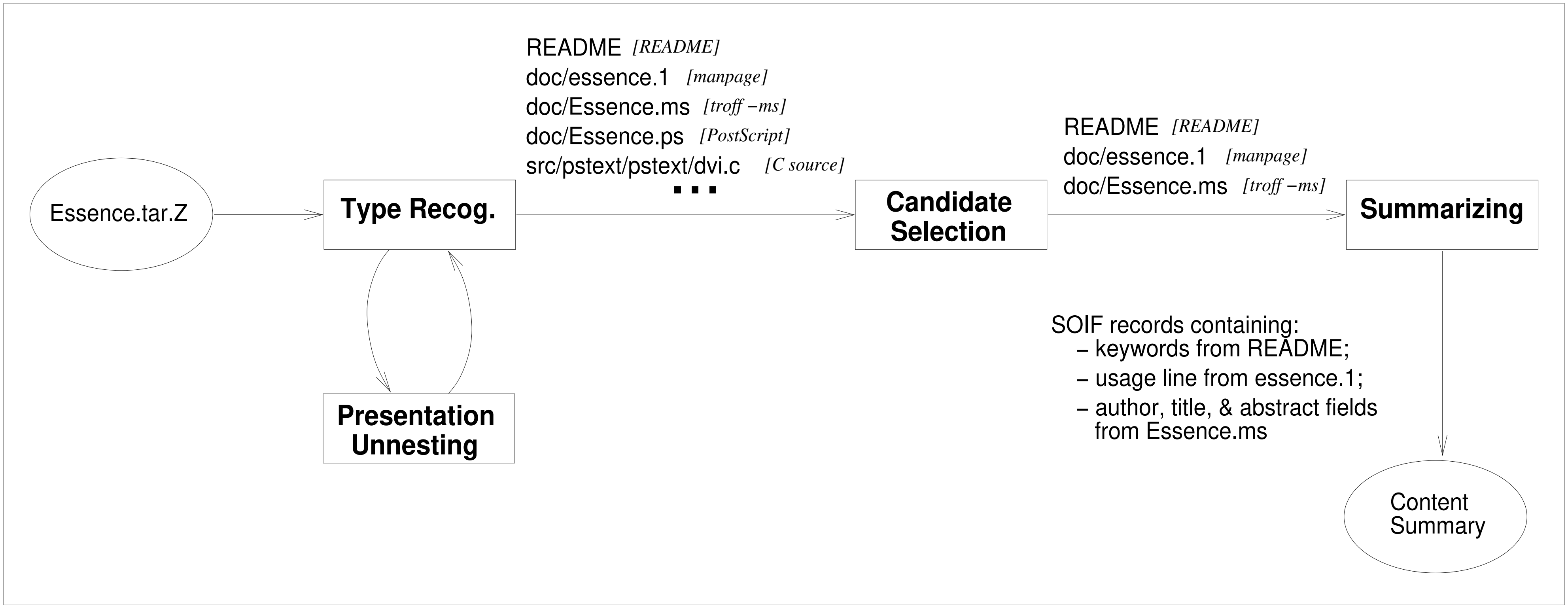

File
Options
Navigate
Annotate
Help
Document Title:
Query Interface to the WWW Home Pages Broker
Document URL:
http://rd.cs.colorado.edu/brokers/www-home-pages/query.html
Query Interface to the WWW Home Pages Broker
Enter the query in the box below. For example, try:
Computer AND Science AND Department
to match the Home Pages with Computer Science Department in the title. Common attributes to use include: Author, Keywords, Title, and URL-References. Look here for help formulating queries; and here for some example WWW Home Page queries. For more information about the Harvest system, look here.
Query:
supreme AND court AND decisions
Press button to submit or reset the query:
Submit
Reset
Query Options:
Case insensitive
Number of errors allowed.
0 Errors
Result Set Options:
Display Matched Lines (from object summaries)
Display Object Descriptions (if available)
Verbose
Back
Forward
Home
Reload
Open...
Save As...
Clone
New Window
Close Window

Summary Object Interchange Format (SOIF):

- structured object summary stream
- arbitrary binary data + extensible fields
- easy interoperation with other tagging standards (IAFA, SGML, etc.)

```
@DELETE { }
@REFRESH { }
@UPDATE {
@FILE { ftp://ecrc.de/pub/ECRC_tech_reports/reports/ECRC-93-10.ps.Z
Time-to-Live{7}:    9676800
Last-Modification-Time{9}:  774988159
Refresh-Rate{7}:    2419200
Gatherer-Name{50}:  Computer Science Technical Reports - Selected Text
Gatherer-Host{21}:  bruno.cs.colorado.edu
Gatherer-Version{3}:    0.3
Type{10}:   Compressed
Update-Time{9}: 774988159
File-Size{6}:   164373
MD5{32}:    43193942d4d53f5a8e4a7b4bcff7a415
Embed<1>-Nested-Filename{13}:   ECRC-93-10.ps
Embed<1>-Type{10}:  PostScript
Embed<1>-File-Size{6}:  428233
Embed<1>-MD5{32}:   84c123582c3d0754a39a78a7e2fb6d23
Embed<1>-Keywords{104}:
technical report ECRC{93{10
Polymorphic Sorts and Types for
Concurrent Functional Programs
Bent Thomsen
}

@FILE { ftp://cml.rice.edu/pub/reports/9404.ps.Z
Time-to-Live{7}:    9676800
Last-Modification-Time{9}:  772872313
Refresh-Rate{7}:    2419200
Gatherer-Name{50}:  Computer Science Technical Reports - Selected Text
Gatherer-Host{22}:  powell.cs.colorado.edu
Gatherer-Version{3}:    1.0
Type{10}:   Compressed
[...rest deleted...]
}
}
```

Customizable Content Extraction (Essence):



Customizable Indexing/Searching

- Generic Broker-Indexer interface accommodates variety of index/search engines w/ single user interface
- Currently: freeWAIS, commercial WAIS, Glimpse, Nebula
- Looking into LSI, Verity, audio/image searching

Demonstration Brokers

Features	Brokers	# Objs	# Sites
• Scalable content indexing • Multiple formats • Gathering distribution	CS Technical Reports	24,200	290
• Incorp. manual indexing info • Support for other standards • Structured Indexing	PC Software Packages	28,600	6
	Linux Software Maps	1,300	84
	IAFA	-	-
• Focused global content index • Structured Indexing	WWW Home Pages	17,600	6,240
• Topic-specific indexing • Broker cascading	NIDR	1,500	83
	NIDR-Document	1,300	70
	NIDR-Software	200	24
• Customizable Indexing • Structured Indexing	Sante Fe Institute Papers	32	4
	AT&T 1-800 phone #s	4,100	1
• Locate Harvest servers • Structured Indexing	Harvest Server Registry	20	3

Implementing Other Systems with Harvest

Archie, Veronica, WWWW, etc.
Gatherer configuration + Essence extraction script

Content Router
WAIS enumerator + Essence extraction script

WAIS
Essence full text extraction + ranking search engine

WebCrawler
Gatherer configuration + Essence full text extraction

WHOIS++
Gatherer configuration + Essence extraction script for centroids. Query front-end to search SOIF records

Measurements

Server load:

gathering via FS + serving single stream → 25,000x savings *per indexer* (average)

Network traffic:

pre-summarized + compressed → 59x savings (not counting savings from incremental updates)

Indexing space:

Allowing compressed data + summarizing (+ space effic. index) → 2-3% of WAIS index space

---

AT&T 1-800 Example:

- 10 hours to gather 4,100 HTML files (20 MB)
- 3 min. to transmit compressed 2.9 MB summary stream

CS Tech Reports Example:

- Using WAIS: 4.5 GB provider data + 4.5 GB index
- Using Harvest: 157 MB content summaries + 118 MB

Demos, papers, software, documentation, email contacts:

http://harvest.cs.colorado.edu/

HARVEST

EFFECTIVE USE OF INTERNET INFORMATION