



connecting people with information

Verity Information Server

Training Guide



Education Department
892 Ross Drive
Sunnyvale, CA 94089
(408) 541-1500

all rights reserved



Basic Collection Building

- ◆ What is a Collection?
- ◆ Tools for Building Collections: Utilities & Spiders
- ◆ The Collection Servicer Utility
- ◆ Managing Your Collections
- ◆ Adding Topics and Indexing Against Collections
- ◆ Working with Utility Programs
- ◆ Practice Lab

What is a Collection?

- ◆ A Verity Collection is a series of indexes which work together to enable the searching of documents
- ◆ Quality searching requires
 - Access to attributes about the document
 - ◆ Title <contains> Verity
 - ◆ Date > 1/1/98
 - Limiting to zones in the document
 - ◆ Verity <in> Title
 - ◆ Danger <in> Subject
 - ◆ (danger, caution, warning) <IN>(h1, h2, title)
 - Proximity information on words in the document
 - ◆ Verity <sentence> new products
 - ◆ Printer <near>/4 problems
 - ◆ White House
- ◆ Verity's indexing utilities automatically capture values for these types of searches and write this data to a collection

How Are Values Populated?

- ◆ Each collection has a style directory that contains the rules for populating the collection's fields, zones and word indexes
 - The indexer “senses” each document's type and load rules for processing that document
 - Standard fields and zones (Title, Date, Keywords, etc.) for various document types are pre-defined by Verity and are populated by automatic filtering utilities
 - You can define custom fields by modifying your style directory
 - ◆ Meta-tags in HTML documents are very easy to use

```
<META name="Dept" content="Sales">
```
 - ◆ Document properties can be populated and extracted by the indexer
 - ◆ A “bulk-submit” file can be created which is handed to the indexer with all field values already defined

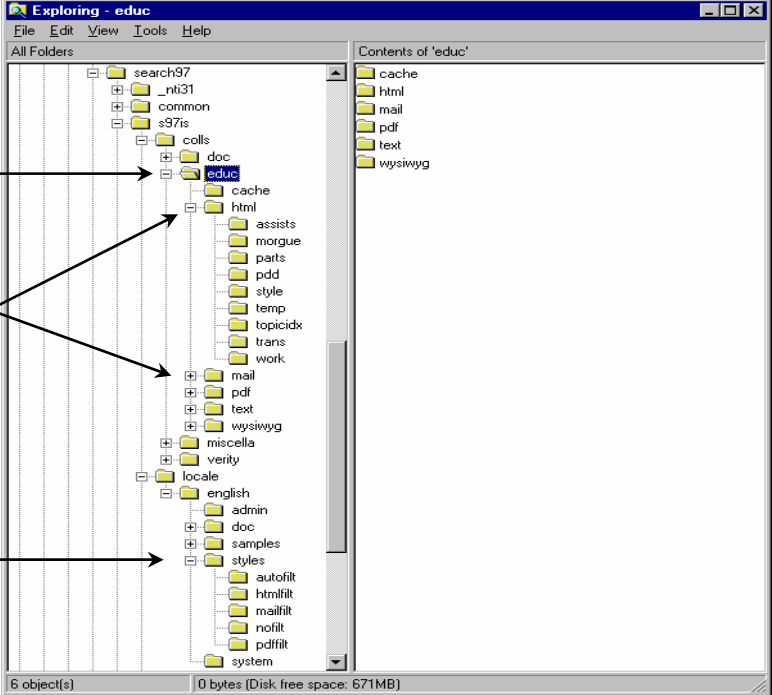


Document Indexes

- ◆ Actual document information is stored in the parts (partitions) directory
 - The “.ddd” stores attribute and location information
 - The “.did” stores word information
- ◆ Verity provides utilities you can use to better understand the information contained in your collection
 - **Browse** allows you to view the values for each of the attributes captured
 - **Didump** allows you to view words in the word index
- ◆ The Information Server also comes with **RCVDK**, a lightweight command line retrieval client which allows you to attach to one or more collections and perform full text searching, viewing of documents and displaying of results
- ◆ Your workbook includes information about these utilities and an exercise to try each of them

Collection Structure

- ◆ The collection directory contains sub-directories that maintain information about:
 - How the collection is to be created and managed (style directory)
 - Data captured from documents like fields, zones, words (parts directory)



virtual collection → educ

physical collections → cache, html, assists, parts, pdd, style

style directories → style

assists

- 00000000.abt
- 00000000.ngm
- 00000000.wld

parts

- 00000001.ddd
- 00000001.did

pdd

- 00000000.pdd

style

- style.ddd
- style.dft
- style.did
- style.ngm
- style.pdd
- style.prm
- style.sid
- style.ufl
- style.vgw
- style.wld

trans

- data.trn

LAST LOGCHECK "-1347706221"
 LAST CLEAN "-1347706221"
 LAST OPTIMIZE "0"
 LAST MAINTENANCE "-1347706207"



How Indexing Works

- ◆ You provide 4 important pieces of information to begin the creation of a collection:
 - Which utility you will use (vspider or mkvdk)
 - Which documents you wish to include (starting directory, URL or specific file names)
 - What you will call your collection and where it will be located
 - Which style files you will use (path to the top-level style directory)

```
vspider -start http://www.yoursite.com  
-collection d:\webcenter\colls\mycoll.clm  
-style d:\verity\s97is\locale\english\styles
```

*commands are entered on one line
there are many additional options you can add*

Accessing Documents

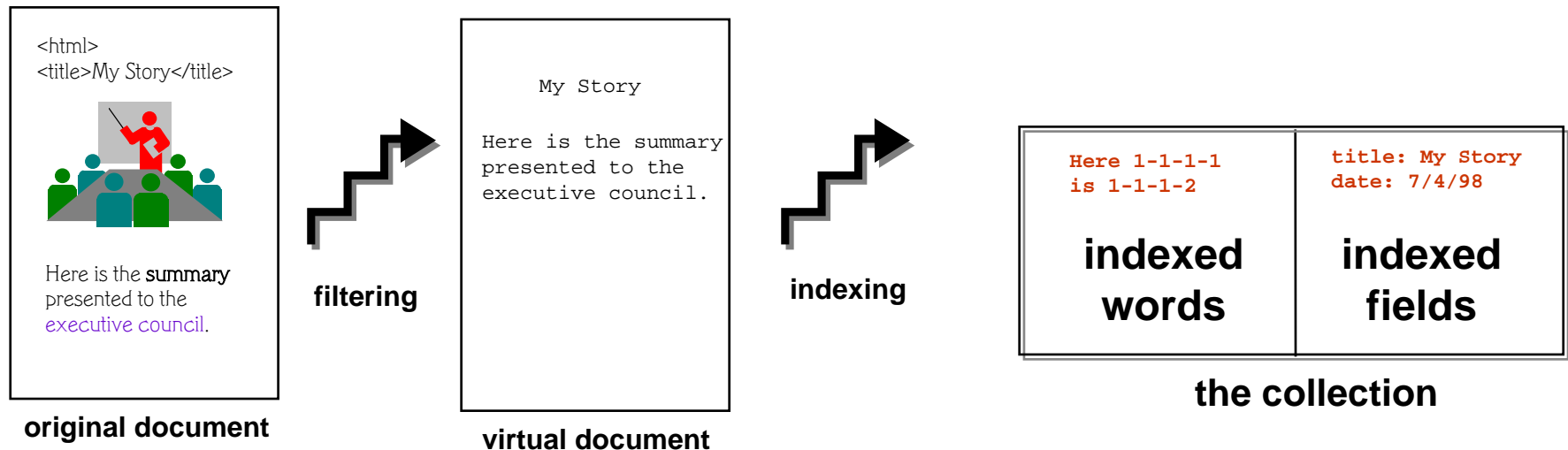
- ◆ The Verity indexing engine uses a gateway to access document files or other repositories of data (on the web, file system or in a database)
 - Default is file system (local host or mapped network drives)
 - Verity provides pre-defined gateways for access through HTTP servers and databases. Custom gateways can be defined.



- ◆ Gateways are specified in the style.vgw (Verity GateWay file). If this file is not found, the default gateway is used.
- ◆ Gateway fields are captured (vgw_url, mime-type, modified date, vgw_odbc: database fields)

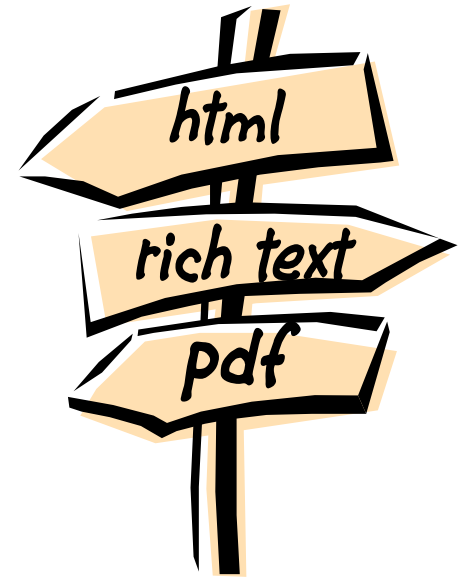
Filtering Documents

- ◆ Each document headed for the collection goes through a temporary conversion process to create a **virtual document** of indexable text
- ◆ An auto-recognition program identifies the type of document to be filtered and calls on one of the helper sub-filters to handle the processing of the selected document
 - Handles characters or binary data as appropriate
 - Removes general formatting tags
 - Creates word, punctuation, markup and field tokens



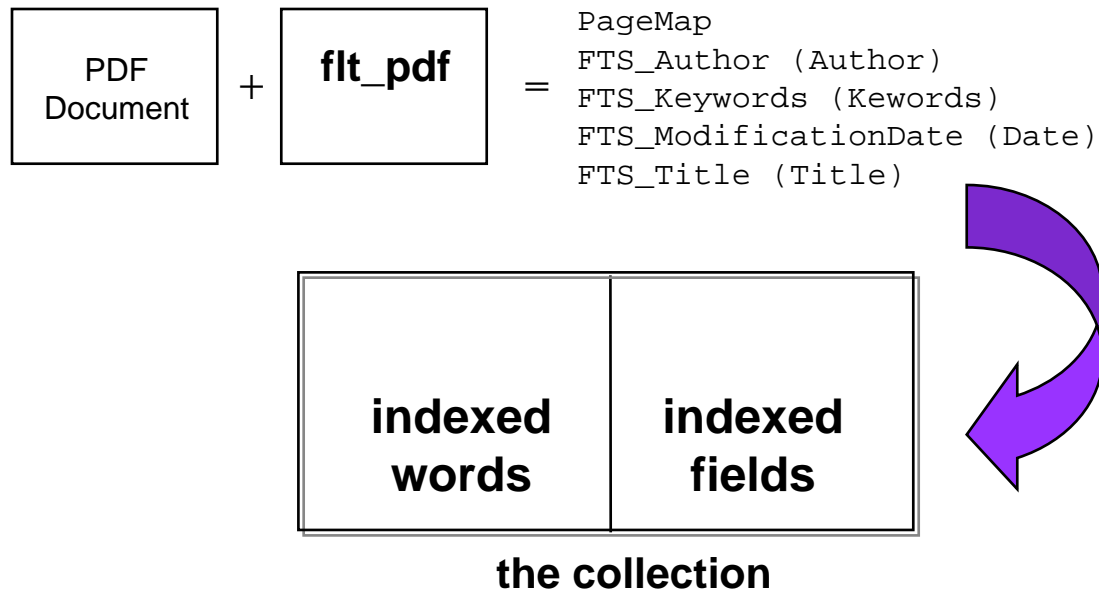
Helper Sub-Filters

- ◆ There are 3 helper sub-filters:
 - **flt_zon** is a Verity filter with options for html, email and news
 - **flt_kv** is a kit of KeyView filters enabling indexing of more than 45 of the most popular document formats (word processing, spreadsheets, presentations)
 - **flt_pdf** is a Verity filter designed to handle PDF documents
- ◆ The number of fields and content varies based on the type of filter that is used



Indexing Documents

- ◆ Document attributes (like title and author) are captured in a field index. They can be populated by gateways, document filters or bulk-submit files.
- ◆ Each document has a record in the index and the fields associated with it are determined by the style files that were used to create the collection
 - Verity identifies key fields for various document types
 - You can customize style files to add additional fields





Browsing Collection Contents

browse - Verity, Inc. Version 2.2.2 (_nti31, Sep 03 1997)

BROWSE OPTIONS

- ?) help
 - q) quit
 - c) Number of entries in field
 - _) Toggle viewing fields beginning with '_'
 - v) Toggle viewing selected fields
 - ##) Display all fields in specified record number
- Dispatch/Compound field options:
- n) No dispatch
 - d) Dispatch
 - s) Dispatch as stream

menu options

Action (? for help): Record number: 0

0	_DDFLAG	FIX-unsq (1) = 0x00
1	_DDVALUE	VAR-text (0) =
2	_DDVALUE_OF	FIX-unsq (4) = 0
3	_DDVALUE_SZ	FIX-unsq (2) = 0
4	_DBVERSION	WRM-text (6) = vdk21
5	_DDDSTAMP	FIX-date (4) = 03-Jan-1998 12:25:42 pm
6	_DOCIDX	VAR-text (12) =
7	_PARTDESC	FIX-text (22) = (Verity, Inc. Version 2.2.2) -
8	_FtrCfg	CON-text (3) = TF
9	_SumCfg	CON-text (27) = XS MaxSents 3 MaxBytes 500
10	_SPARE1	FIX-text (15) =
11	_SPARE2	FIX-sign (4) = 0
12	_DOCIDX_OF	FIX-unsq (4) = 32
13	_DOCIDX_SZ	FIX-unsq (2) = 12
14	_STYLE	AUT-text (19) = ../style/style.ddd
15	_DOCID	FIX-unsq (4) = 3
16	_SECURITY	FIX-unsq (4) = 0
17	_INDEX_DATE	FIX-date (4) = 03-Jan-1998 12:25:43 pm
18	_SECURITY_MI	WRM-unsq (4) = 0
19	_SECURITY_MX	WRM-unsq (4) = 0

field types

feature vectors for clustering and stored summaries

when document was indexed

more...



Browsing Collection Contents

```

20 _INDEX_DATE_MI    WRM-date ( 4) = 03-Jan-1998 12:25:43 pm
21 _INDEX_DATE_MX    WRM-date ( 4) = 03-Jan-1998 12:25:43 pm
22 VDKFEATURES       VAR-text (232) =
23 VDKFEATURES_OF    FIX-unsq ( 4) = 539
24 VDKSUMMARY        VAR-text (271) = Beau Geste. Based on Christopher Wren's novel
about the undying devotion shared among three brothers serving in the French Foreign
Legion. Gary Cooper and Ray Milland join a number of to-be-famous character actors in
this exciting adventure movie. Released: 1939 Date: 13 Nov 1993
25 VDKSUMMARY_OF    FIX-unsq ( 4) = 857
26 VdkVgwKey         VAR-text ( 27) = c:\is97\docs\doc3\Rev1.txt
27 VdkVgwKey_IK      FIX-unsq ( 3) = 2
28 VdkVgwKey_MI      WRM-text ( 28) = c:\is97\docs\doc3\MSG17.TXT
29 VdkVgwKey_MX      WRM-text ( 27) = c:\is97\docs\doc3\REV6.TXT
30 VdkVgwKey_OF      FIX-unsq ( 4) = 88
31 VdkVgwKey_SZ      FIX-unsq ( 2) = 27
32 DOC               DSP-text ( -1) = c:\is97\docs\doc3\Rev1.txt
33 DOC_FN            VAR-text ( 27) = c:\is97\docs\doc3\Rev1.txt
34 _CACHE_FN         VAR-text ( 0) =
35 _CACHE_DELETE     FIX-unsq ( 1) = 0
36 _ParentID         VAR-text ( 18) = c:\is97\docs\doc3
37 Title             VAR-text ( 5) = Beau Geste Review
38 Ext               VAR-text ( 4) = Txt
39 Author            VAR-text ( 0) =
40 Subject           VAR-text ( 0) =
41 Keywords          VAR-text ( 0) =
42 Comments          VAR-text ( 0) =
43 Snippet           VAR-text (288) = Beau Geste. Based on Christopher Wren's novel
about the undying devotion shared among three brothers serving in the French Foreign
Legion. Gary Cooper and Ray Milland join a number of to-be-famous character actors in
this exciting adventure movie. Released: 1939 Date: 13 Nov 1993

```

auto-generated summary

primary key

the document locator

Index - Min/Max

Offset - Size

document file name

where document was found

document title as identified

by the filtering helper used

fields built by default

but not populated by filter

first 400 printable characters

more...



Summary vs. Snippet

Different because the summary is limited to only the best two lines

24 VDKSUMMARY VAR-text (240) = This chapter covers basic information about SEARCH'97(TM) Information Server and the SEARCH'97 search technology from Verity© that is integrated in the product. The following subjects are included: Introduction to Verity Search Technology.

43 Snippet VAR-text (399) = Introduction 1 . Introduction. This chapter covers basic information about SEARCH'97(TM) Information Server and the SEARCH'97 search technology from Verity® that is integrated in the product. The following subjects are included: . Introduction to Information Server . New SEARCH'97 Features . Introduction to Verity Search Technology . Copyright © 1997, Verity, Inc. All rig...

Same because the document is so short and summaries was set to a 3 sentence maximum

24 VDKSUMMARY VAR-text (271) = Beau Geste. Based on Christopher Wren's novel about the undying devotion shared among three brothers serving in the French Foreign Legion. Gary Cooper and Ray Milland join a number of to-be-famous character actors in this exciting adventure movie. Released: 1939 Date: 13 Nov 1993

43 Snippet VAR-text (288) =Beau Geste. Based on Christopher Wren's novel about the undying devotion shared among three brothers serving in the French Foreign Legion. Gary Cooper and Ray Milland join a number of to-be-famous character actors in this exciting adventure movie. Released: 1939 Date: 13 Nov 1993

more...



Browsing Collection Contents

if web document, URL here

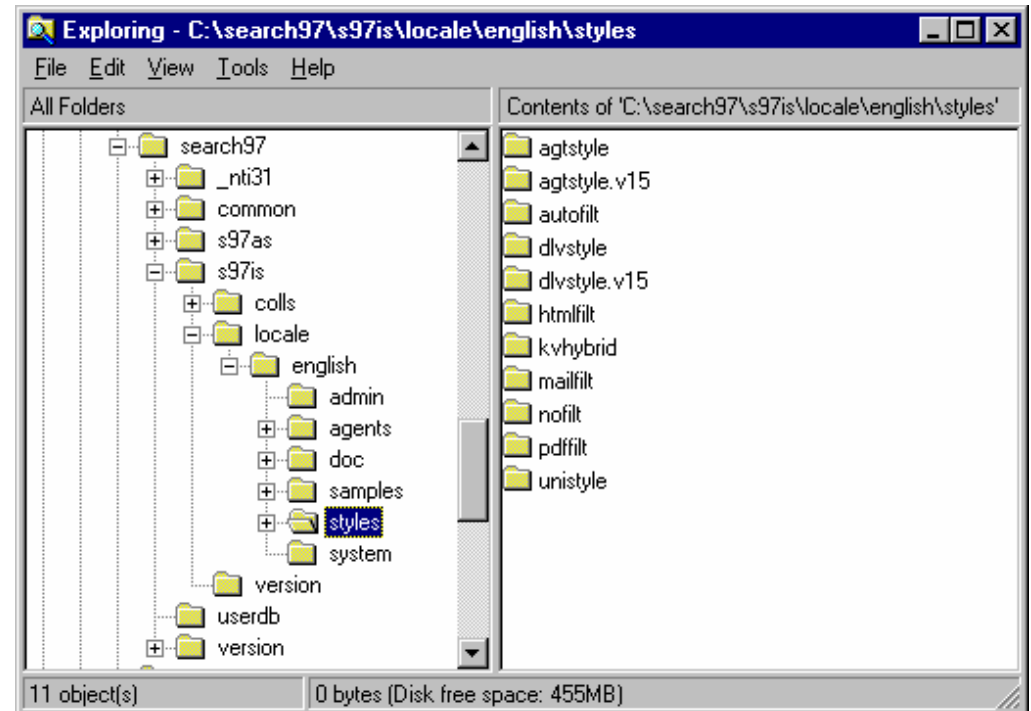
mime type impacts viewing

date information: some as text fields some as Verity-Internal Dates

44	URL	VAR-text (0) =	
45	MIME-Type	VAR-text (11) = text/plain	46 Language
47	Encoding	VAR-text (0) =	
48	_Created	VAR-text (30) = Sun, 29 Dec 1997 23:37:35 GMT	
49	_Modified	VAR-text (30) = Sun, 23 Oct 1997 02:27:44 GMT	
50	Created	FIX-date (4) = 29-Dec-1997 03:37:35 pm	
51	Modified	FIX-date (4) = 22-Oct-1997 06:27:44 pm	
52	Size	FIX-unsg (4) = 313	
53	DOC_OF	FIX-unsg (4) = 0	
54	DOC_SZ	FIX-unsg (4) = 4294967295	
55	DOC_FN_OF	FIX-unsg (4) = 88	
56	DOC_FN_SZ	FIX-unsg (2) = 27	
57	_CACHE_FN_OF	FIX-unsg (4) = 0	
58	_CACHE_FN_SZ	FIX-unsg (2) = 0	
59	_ParentID_OF	FIX-unsg (4) = 2246	
60	_ParentID_SZ	FIX-unsg (2) = 18	
61	Title_OF	FIX-unsg (4) = 2237	
62	Title_SZ	FIX-unsg (2) = 5	
63	Ext_OF	FIX-unsg (4) = 2242	
64	Ext_SZ	FIX-unsg (2) = 4	
65	Author_OF	FIX-unsg (4) = 0	
66	Author_SZ	FIX-unsg (2) = 0	
67	Subject_OF	FIX-unsg (4) = 0	
68	Subject_SZ	FIX-unsg (2) = 0	
69	Keywords_OF	FIX-unsg (4) = 0	
70	Keywords_SZ	FIX-unsg (2) = 0	
71	Comments_OF	FIX-unsg (4) = 0	
72	Comments_SZ	FIX-unsg (2) = 0	
73	Snippet_OF	FIX-unsg (4) = 2264	
74	Snippet_SZ	FIX-unsg (2) = 288	
75	URL_OF	FIX-unsg (4) = 0	
			76 URL_SZ
			FIX-unsg (2) = 0
			77 MIME-Type_OF
			FIX-unsg (4) = 174
			78 MIME-Type_SZ
			FIX-unsg (2) = 11
			79 Language_OF
			FIX-unsg (4) = 0
			80 Language_SZ
			FIX-unsg (2) = 0
			81 Encoding_OF
			FIX-unsg (4) = 0
			82 Encoding_SZ
			FIX-unsg (2) = 0
			83 _Created_OF
			FIX-unsg (4) = 185
			84 _Created_SZ
			FIX-unsg (2) = 30
			85 _Modified_OF
			FIX-unsg (4) = 215
			86 _Modified_SZ
			FIX-unsg (2) = 30

Configuring Collections

- ◆ Style files configure the indexes that store data about documents
 - Choices made in style files direct how indexing utilities will create and maintain collections
 - Information Server ships with a directory of styles for your use (default is shown below)
- ◆ Advanced Collection Building on day 3 covers style files for all document types and features, in greater detail





Indexing and Logging

- ◆ You can build or add to collections from either the GUI Indexer or the command line indexing utilities (vspider or mkvdk)
- ◆ Log files are automatically created and you can set the level of feedback you desire

— For the GUI indexer set this in the inetsrch.ini file

```
LogLevel = Verbose    (less information)
LogLevel = Debug      (more information)
LogLevel = Trace      (most information)
```

— For vspider, set this on the command line

```
-verbose    -debug    -trace
```



Skipping Documents

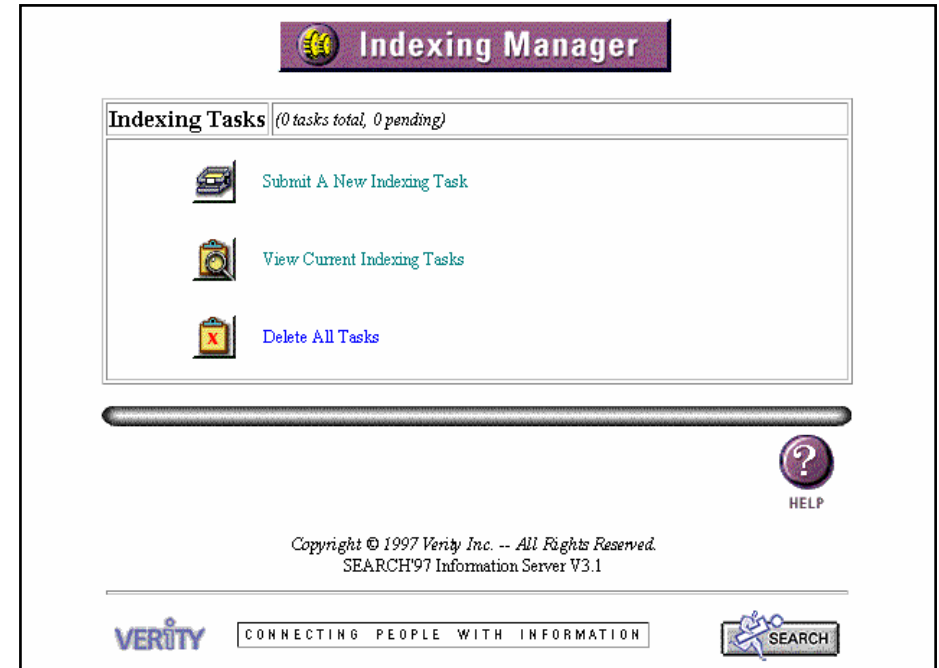
- ◆ You may see messages about skipped documents in the GUI Indexing Manager or in your log files
- ◆ Skipping messages: “Skipping key because of...”

HOST	Document is not on the host you are indexing
DOMAIN	Document is not in the domain you are indexing
DATE	Document is up-to-date in the index and does not require reindexing
VDKVGWKEY	Path to the document does not match your -include statement or does not match your -exclude statement
MIME-TYPE	The document is of a mime type you are not indexing

- ◆ Also skips if off-site, forbidden access (401) or just doesn't exist (404)

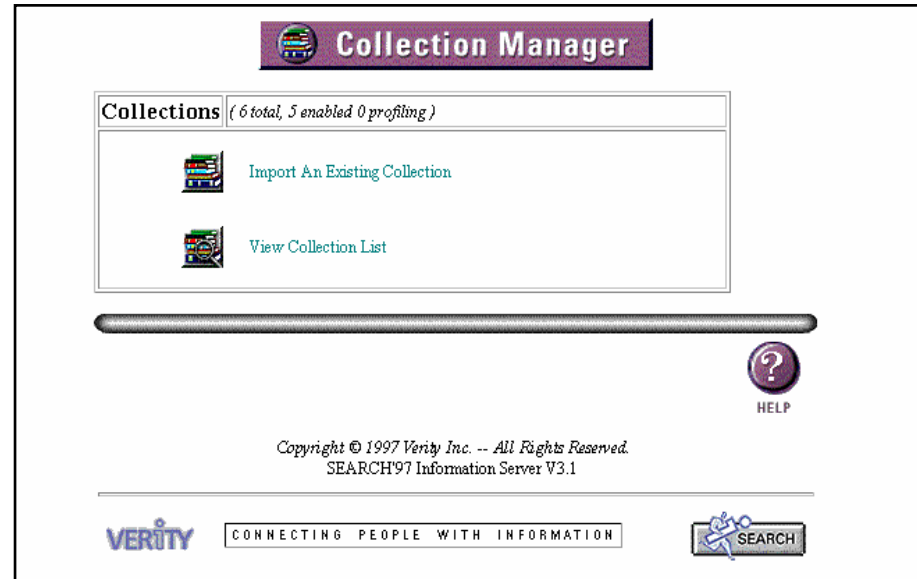
The Graphical Indexing Manager

- ◆ The Indexing Manager organizes features related to indexing specific sites
 - Simple mode only requires a name, description and path
 - Advanced mode filters for MIME types, include or exclude patterns (file or domain names), and sets proxy information at the collection level
- ◆ Indexing tasks can be defined and submitted for initial creation and then maintained as current tasks for resubmission



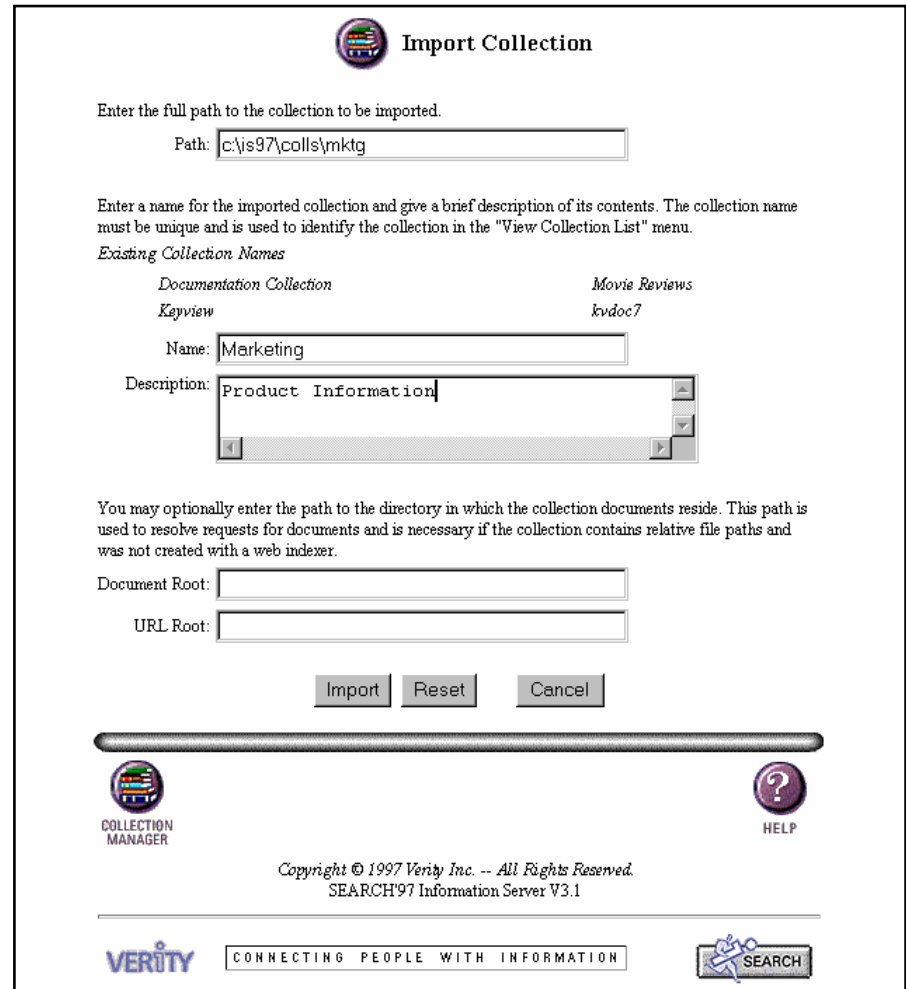
The Graphical Collection Manager

- ◆ The Collection Manager allows you to
 - Work with the current set of collections on your server
 - Import a new collection created or used by another Verity application
 - View information about the state and contents of each collection



Importing Collections

- ◆ Collections built with previous versions of Verity's web indexers import easily
- ◆ If you wish to take advantage of
 - The latest filtering options provided by KeyView filters
 - Performance improvements provided by the new vspideryou will need to rebuild your collections



Import Collection

Enter the full path to the collection to be imported.

Path:

Enter a name for the imported collection and give a brief description of its contents. The collection name must be unique and is used to identify the collection in the "View Collection List" menu.

Existing Collection Names

<i>Documentation Collection</i>	<i>Movie Reviews</i>
<i>Keyview</i>	<i>kvdoc7</i>



Name:

Description:



You may optionally enter the path to the directory in which the collection documents reside. This path is used to resolve requests for documents and is necessary if the collection contains relative file paths and was not created with a web indexer.

Document Root:

URL Root:

 COLLECTION MANAGER  HELP

Copyright © 1997 Verity Inc. -- All Rights Reserved.
SEARCH97 Information Server V3.1

Using the Collection List

- ◆ As you build a number of collections, you will work with them on the Collection List
 - Allows you to easily enable and disable searching
 - “Edit” takes you to Collection Properties
 - You can remove collections from the list, but this does not delete them from the system



The screenshot shows the 'Collection List' interface. At the top, there are 'Enable All' and 'Disable All' buttons, and a red 'X' icon. Below these are four collection entries, each with a green status indicator and the word 'Enabled' below it:

- Documentation Collection
- Movie Reviews
- Keyview
- kvdoc7

Each entry has a set of control icons: a red circle with a white dot (disable), a green circle with a white dot (enable), a blue circle with a white dot (edit), and a red 'X' (remove). Additionally, there are icons for 'P' (disable Profiling) and 'P' (enable Profiling).

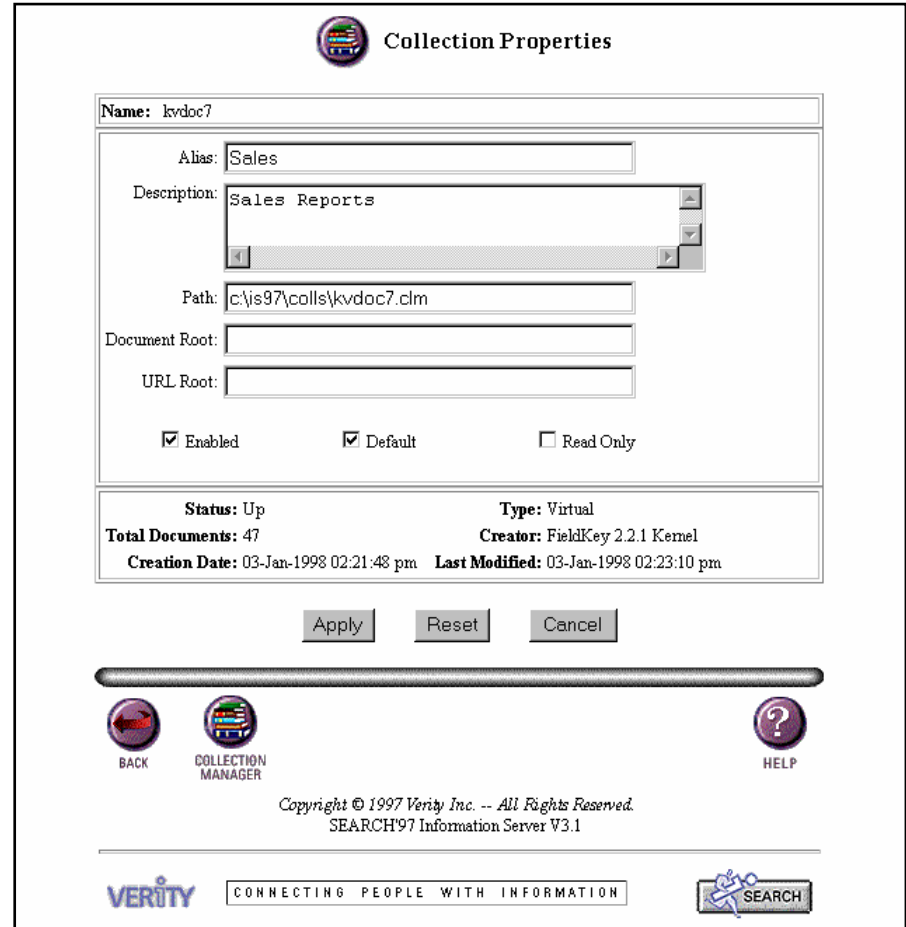
A legend below the entries explains the icons:

- Red circle with white dot = disable
- Green circle with white dot = enable
- Blue circle with white dot = edit
- Red X = remove
- P (with red border) = disable Profiling
- P (with green border) = enable Profiling

At the bottom of the interface, there are buttons for 'COLLECTION MANAGER' and 'HELP', a copyright notice: 'Copyright © 1997 Verity Inc. -- All Rights Reserved. SEARCH97 Information Server V3.1', and the 'VERITY' logo and 'SEARCH' button.

Collection Properties

- ◆ Collection Properties provides information about the current status of the collection
- ◆ It accepts edits for collection alias and description



Collection Properties

Name: kvdoc7

Alias: Sales

Description: Sales Reports

Path: c:\js97\colls\kvdoc7.clm

Document Root:

URL Root:

Enabled Default Read Only

Status: Up Type: Virtual

Total Documents: 47 Creator: FieldKey 2.2.1 Kernel

Creation Date: 03-Jan-1998 02:21:48 pm Last Modified: 03-Jan-1998 02:23:10 pm

Apply Reset Cancel

BACK COLLECTION MANAGER HELP

Copyright © 1997 Verity Inc. -- All Rights Reserved.
SEARCH'97 Information Server V3.1

VERITY CONNECTING PEOPLE WITH INFORMATION SEARCH



The Command-Line Verity Spider

- ◆ The command-line spider (**vspider**) provides additional indexing options and greater control than available through the GUI spider found in the Information Server Indexing Manager
 - Walks the file system or crawls one or more web sites
 - Installed with Information Server by default to allow web spidering for the local host and the file system

```
vspider -collection mycoll.clm  
-style \verity\s97is\locale\english\styles  
-start http://www.yoursite.com/
```

```
vspider -collection mycoll.clm  
-style \verity\s97is\locale\english\styles  
-start c:\webctr\docs\
```

```
vspider -collection mycoll.clm  
-style \verity\s97is\locale\english\styles  
-start http://www.yoursite.com/  
-exclude *secrets* -domain verity.com -proxy proxysrvr:8010
```


Features of Verity Spider

- ◆ This spider allows you to:
 - Specify multiple “start” URLs or starting directories
 - Force the re-parsing of all documents in the collection
 - Limit indexing to a particular domain or host
 - Include or exclude files matching regular expressions or mime types
 - Set a maximum size for documents to be indexed
 - Disable the following of links
 - Disable updating of documents already in the collection
 - Specify topics to be used when indexing the collection
 - Specify logging message levels (verbose, debug, trace)
 - Read command-line syntax for vspider from a file
 - Specify the import date format to use
 - Specify the number of URLs to be streamed simultaneously (default is 10)
 - Disable DNS look ups (for faster spidering)
- ◆ A complete list of options are available in your workbook



Verity Spider Licensing

- ◆ Licensing options include
 - File walking through any files available on the network (automatically included with any of the other options)
 - Web spidering of local host for files and links contained on the Information Server's host machine (default)
 - Web spidering of default domain for the local host
 - Web spidering of remote sites
- ◆ Host refers to physical machine - these are different hosts
 - <http://www.verity.com>
 - <http://uk.www.verity.com>
 - <http://web.verity.com>
- ◆ Domain refers to the last two entries of the DNS name, regardless of location

Authentication

- ◆ Some sites secure documents by requiring authentication through their web server
- ◆ You can use the `-auth` option to specify the name of your authentication file. This file includes the server name, realm, username and password:

```
#server          #realm #username #password
www.verity.com  field  Alladin  "sesame"
```

- ◆ Storing the username and password allows users to see all documents within the authenticated path
- ◆ Users will be able to view highlights within the retrieved documents

```
vspider -collection mycoll.clm
        -style /verity/s97is/locale/english/styles
        -start http://www.yoursite.com/
        -auth authfile.txt
```

Recognized MIME Types

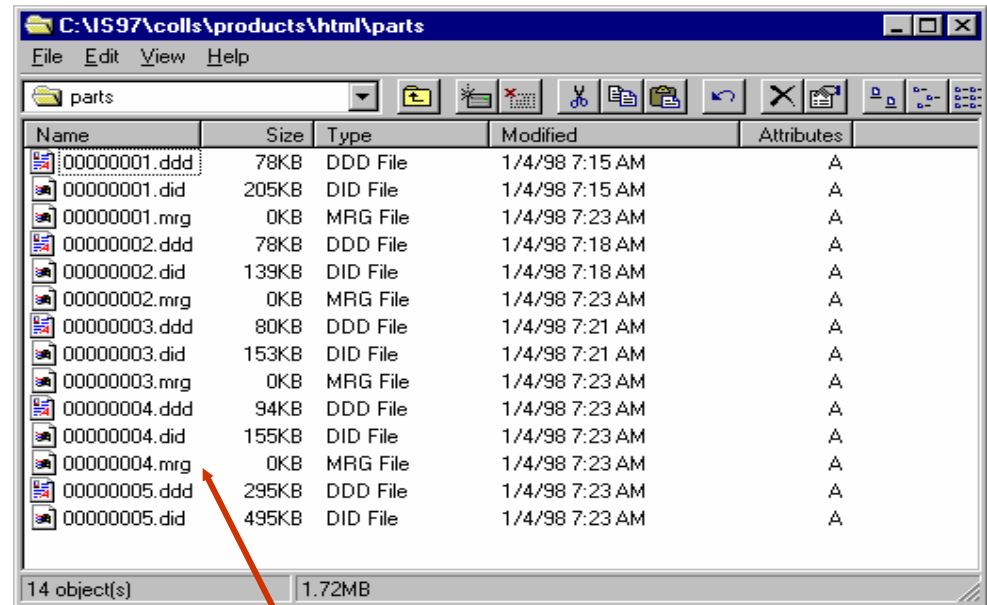
- ◆ The vspider utility recognizes these MIME types when Universal filtering or matching for -include and -exclude options

Format	MIME Type	Filter
HTML	text/html	zone filter
SGML	text/sgml	zone filter
News	message/news	zone filter
Email	message/rfc822	zone filter
ASCII	text/*	(no filter)
RTF	application/rtf	Keyview
PDF	application/pdf	PDF filter
GIF	image/gif	(no filter)
MS Word	application/msword	Keyview
MS Excel	application/x-ms-excel	Keyview
MS Powerpoint	application/x-ms-powerpoint	Keyview
WordPerfect	application/wordperfect5.1 application/x-corel-wordperfect	Keyview
MS Works	application/x-ms-works	Keyview
MS Project	application/x-ms-project	Keyview
Lotus AMI Pro	application/x-lotus-amipro	Keyview
Lotus 1-2-3	application/x-lotus-123	Keyview

- ◆ Other MIME types can be filtered by invoking the appropriate filter in style.uni

Maintaining Your Collections

- ◆ As documents are added to a collection, multiple instances of the .ddd and .did are created
 - Multiple indexing tasks are submitted
 - Large numbers of files or long indexing processes create “chunks” at a time
- ◆ Merging is the process of taking several small partitions and creating a larger single file that is faster to read
- ◆ Housekeeping is the process of cleaning up what is no longer needed



indicates this partition has been merged and will be deleted



Collection Performance

- ◆ Collections begin merging when the number of partitions exceeds 4 (4 → 1)
- ◆ Housekeeping occurs when
 - An indexing process has been running for more than 5 minutes
 - A scheduled service is run by the Collection Servicer utility
 - An mkvdk instruction to service the collection has been issued
 - ◆ Housekeep deletes unneeded files
 - ◆ Optimize enables background optimization
 - ◆ Data prep handles any outstanding work including optimization and housekeeping

```
mkvdk -collection c:\is97\colls\products.clm -servlev housekeep
```

```
mkvdk -collection c:\is97\colls\products.clm -servlev optimize
```

```
mkvdk -collection c:\is97\colls\products.clm -servlev dataprep
```

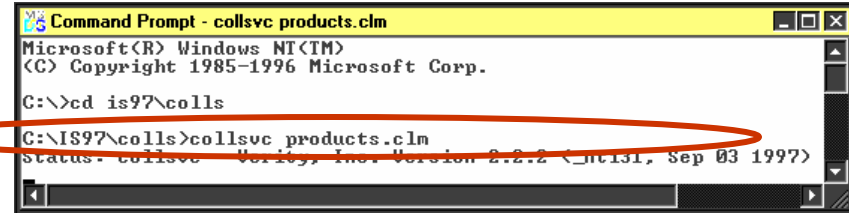


The Collection Servicer Utility

- ◆ Command-line utility for
 - Performing maintenance on collections
 - Offloading processing from vspider
- ◆ Performs any combination of these tasks
 - Insertion of new documents (indexing)
 - Collection optimization (merging)
 - Periodic deletion of document references (housekeeping)
 - Periodic recovery of no-longer-needed disk space (squeezing)
- ◆ Once a collsvc process is executed from the command-line, it continues to run
 - Scheduling parameters for maintenance actions
 - Uses resources as needed
 - On UNIX set a shared library environment variable

Using collsvc

- ◆ Issue the command to service your collection
 - Add document deletion information (delete query and delete schedule)
 - Add partition squeeze information to recapture space after a delete (squeeze schedule)
 - Virtual collections don't allow deletes or squeezes so you will have to set a collsvc process for each of the real collections



```

Command Prompt - collsvc products.clm
Microsoft(R) Windows NT(TM)
(C) Copyright 1985-1996 Microsoft Corp.
C:\>cd is97\colls
C:\IS97\colls>collsvc products.clm
Status: collsvc Verity, Inc. Version 2.2.2 (Jul151, Sep 03 1997)
    
```

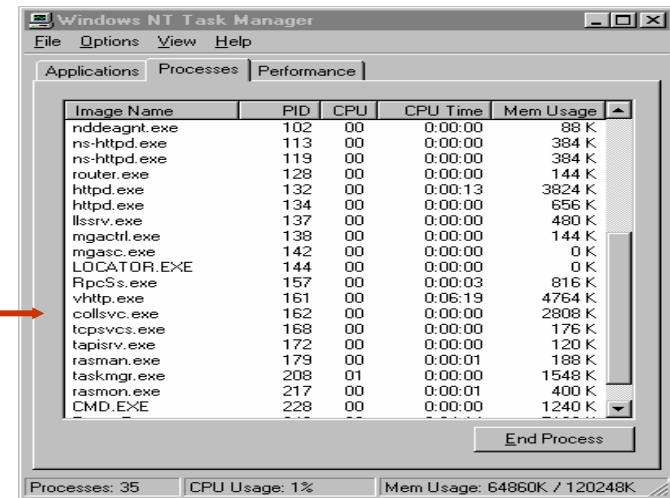


Image Name	PID	CPU	CPU Time	Mem Usage
nddeagnt.exe	102	00	0:00:00	88 K
ns-httpd.exe	113	00	0:00:00	384 K
ns-httpd.exe	119	00	0:00:00	384 K
router.exe	128	00	0:00:00	144 K
httpd.exe	132	00	0:00:13	3824 K
httpd.exe	134	00	0:00:00	656 K
llsrv.exe	137	00	0:00:00	480 K
mgactrl.exe	138	00	0:00:00	144 K
mgasc.exe	142	00	0:00:00	0 K
LOCATOR.EXE	144	00	0:00:00	0 K
RpcSs.exe	157	00	0:00:03	816 K
vhttp.exe	161	00	0:06:19	4764 K
collsvc.exe	162	00	0:00:00	2808 K
tcpvcs.exe	168	00	0:00:00	176 K
tapisrv.exe	172	00	0:00:00	120 K
rasman.exe	179	00	0:00:01	188 K
taskmgr.exe	208	01	0:00:00	1548 K
rasmon.exe	217	00	0:00:01	400 K
CMD.EXE	228	00	0:00:00	1240 K

```

Sun 04-Jan-98 09:48:14 AM - 162 - Status: Opening collection
vdksvc.exe (Sun Jan 04 09:48:15 1998): Status E1-0103
(Vdksvc: Coll): Commencing servicing collectionvdksvc.exe
(Sun Jan 04 09:48:16 1998): Status E1-0007 (Vdksvc): vdksvc
shutting down98 09:48:15 AM - 220 - Status: vdksvc shutting
down
    
```

collsvc.log

The browse Utility

- ◆ The browse utility provides information regarding the fields and other document attributes in a single partition
- ◆ Output can be viewed on screen or redirected to a file
`browse 00000001.ddd > myfile.txt`
- ◆ It is helpful to understand menu options, as you have to simulate these to redirect output. Your workbook takes you through this process.

```
browse.exe - Verity, Inc. Version 2.2.0 (_nti31, May 2 1997)
```

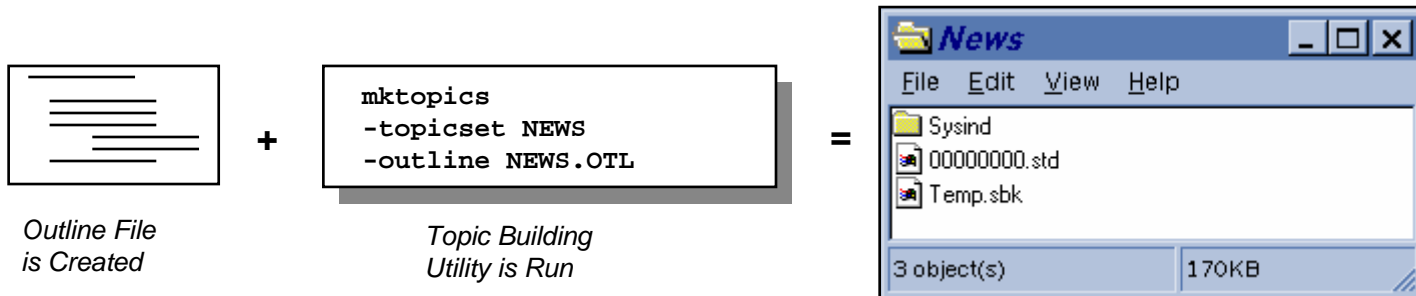
```
BROWSE OPTIONS
```

```
? ) help  
q ) quit  
c ) Number of entries in field  
_ ) Toggle viewing fields beginning with '_'  
v ) Toggle viewing selected fields  
##) Display all fields in specified record number  
Dispatch/Compound field options:  
n ) No dispatch  
d ) Dispatch  
s ) Dispatch as stream
```

```
Action (? for help): Record number: 0
```

The mktopics Utility

- ◆ You create topic definitions for a topic set in a topic outline file
- ◆ The mktopics utility creates a compiled topic set from an outline file
 - Syntax is checked. If an error is detected the utility provides you with information to correct the error.
 - If no errors are found, the mktopics build the topic set in the directory specified
 - The maximum number of topics is 500,000 and the maximum number of topic links is 800,000
 - Create a new topicset: `mktopics -topicset news -outline news.otl`





Indexing Collections with Topics

- ◆ There are two ways to work with a compiled topic set:
 - Add it to the Information Server to help users write better queries
 - Create a special topic index for your collection to speed retrievals (33x faster)
- ◆ Enter the path to your topicset in the “common” section of the inetsrch.ini file
- ◆ Topics will be substituted for the words entered by users. You can also use these topics on your search forms.

```
Common  
TopicSet=c:\is97\topics\news
```

inetsrch.ini

searches expand automatically
against the word index
sports = sports, baseball, soccer, football

```
vspider -collection c:\is97\colls\mycoll.clm  
-style c:\verity\s97is\locale\english\styles  
-start http://www.yoursite.com/  
-topicset c:\is97\topics\news
```

pre-built answer sets
instantly presented
sports: 7,52,133,619,705

The didump Utility

- ◆ The didump utility provides information regarding the occurrences of words in a single partition

- ◆ Output can be viewed on screen or redirected to a file

```
didump 00000000.did > myfile.txt
```

- ◆ Details can be obtained on all words or a single word

```
didump [-v] [-pattern pattern] partition.did
```

```
-v          produce verbose listing per word
```

```
-pattern    word or regular expression to display
```

Sample Output

Word	Size	Doc	Occurrences
A	371	42	81
a	4382	66	1394



RCVDK

- ◆ RCVDK is a great little command line retrieval client that will allow you to quickly test your collections.

```
Available commands:

search          s          Search documents
results        r          Display search results
clusters       c          Display clustered search results
view           v          View document
summarize      z          Summarize documents
attach         a          Attach to one or more collections
detach         d          Detach from one or more collections
quit           q          Leave application
about          ?          Display TDK "about" info
help           ?          Display help (use help help for details)
expert         x          Toggle expert mode on/off
```

```
rcvdk collection-name
rc v2.2
Attaching to collection: collection-name
Successfully attached to 1 collection
Type 'help' for a list of commands
RC>
```



Practice Lab

- ◆ Please complete the exercises for Practice Lab #2 in your student workbook
 - These exercises will give you a chance to gain an understanding of
 - ◆ Basic structure and contents of a collection
 - ◆ Collection management features
 - ◆ Collection building with vspider
 - ◆ Using mkvdk to optimize collections
 - ◆ Adding topics to your application
 - ◆ Using collection information utilities



connecting people with information

Enabling Search at the Server

- ◆ Exploring the Query Language
- ◆ Search Tips Online Guide
- ◆ Search Form Alternatives
- ◆ Practice Lab

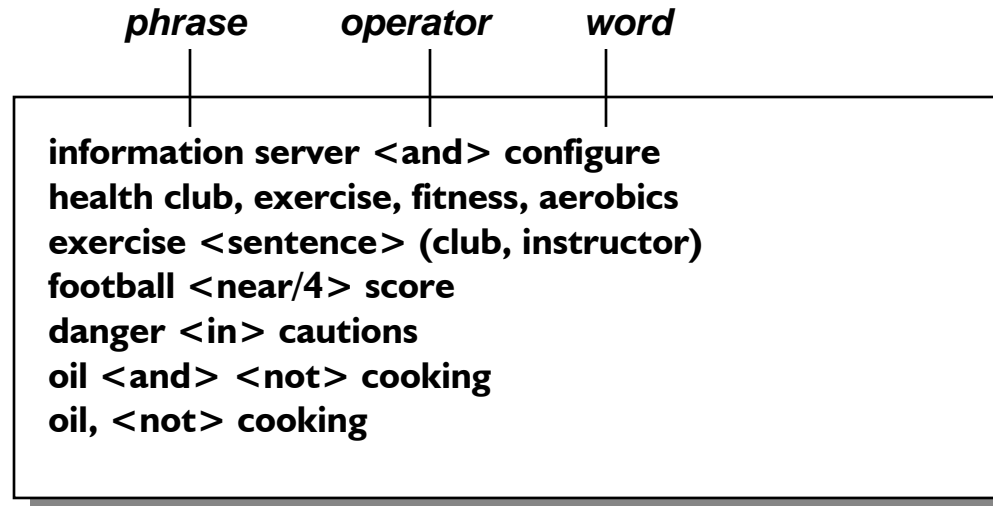


Exploring the Query Language

- ◆ The value of a query language is seen in two areas
 - How easy is it for a novice to ask for information and get good results?
 - How rich is the language for knowledge workers to precisely target specific results?
- ◆ For the novice, the query language provides defaults
 - Standard variant endings
 - Density tie-breakers
 - Accrue on groups of words
 - Ranking by score with highest first
- ◆ For the knowledge-worker, the query language provides
 - A rich set of operators and modifiers
 - Parenthetic representation of complex ideas
 - Weighting of terms or groups of terms
 - Topics

What is a Query?

- ◆ A *query* is the criteria you provide for performing a search
- ◆ When you create queries, you can combine words, phrases, fields and topics with operators and modifiers to direct which documents will be selected and how they will be ordered on the results list





Query Components

- ◆ Queries can include any of these components
 - **WORDS** (in your documents and stored in the word index)
 - **TOPICS** (pre-defined groups of words and information about how the words relate to each other)
 - **FIELD VALUES** (attributes about the documents captured as they are added to the collection and stored in the field index)
 - **OPERATORS** (provided by Verity to help you specify words should be searched for and how results should be evaluated)

Syntax Alternatives

- ◆ When you use **simple syntax** (the default), the query is interpreted with a broad focus
 - Searches case-insensitive
 - Applies STEM operator to search words, selecting variant endings
 - Applies the MANY modifier for search words to score documents higher based on word density
 - Automatically interprets words that are topic names as topics
 - Activates the ACCRUE operator at the parent level to specify selection of any of the words entered, but higher scoring of the document for additional occurrences of unique words
- ◆ When you use **explicit syntax**, you instruct the engine about how the search is to be handled. There are shortcuts for some explicit syntax operators and modifiers:
 - <WORD>film or “film”
 - <STEM> film or ‘film’
 - <SOUNDEX> @film@



Operator Classes

- ◆ **Evidence Operators** search for words and can expand into a list of related search words, depending on the operator selected
- ◆ **Proximity Operators** are used with groups of words to define how closely they are related to each other
- ◆ **Concept Operators** combine the meaning of search words to identify a concept in a document
- ◆ **Relational Operators** are used with fields
- ◆ **Zone Operator** is used with HTML zones
- ◆ **Boolean Operators** are used with topics and words to retrieve the elements you describe without operator precedence conflicts
- ◆ **Natural Language Operators** use natural language syntax to a search

more...

Evidence Operators

- ◆ **Evidence Operators** search for words and can expand into a list of related search words, depending on the operator selected

Operator	Shortcut Rule	
<word> film	“film”	must locate an exact match on the word as entered (no variant endings to be included)
<stem> film	‘film’	must locate a match on the root of the word and includes all standard variant endings (filming, filmed, films)
<thesaurus> film		must search for all synonyms listed in the embedded thesaurus for this word
<wildcard> tech*		must match the character string entered with selected variables: <i>fil*</i> <i>substitutes any characters for *</i> <i>fil?</i> <i>substitutes single letter for ?</i>

Proximity Operators

- ◆ **Proximity Operators** are used with groups of words to define how closely they are related to each other. Proximity refinement often improves query results dramatically.

Operator	Shortcut	Rule
<phrase> nice job	nice job	must locate words in the order defined
new <sentence> film		must locate words in the same sentence (any order)
hit <paragraph> film		must locate words in the same paragraph (any order)
weather <near> report		must locate words within 1000 words of each other. Reflects proximity by score, with closest achieving highest score)
football <near/5>score		must locate words within the number of words specified by /n
danger <in> title		locates documents containing values in specific regions (identified during indexing by the zone filter)



Concept Operators

- ◆ **Concept Operators** combine the meaning of search words to identify a concept in a document
- ◆ **Boolean Operators** are used with topics and words to retrieve the elements you describe without operator precedence conflicts (hit or miss)

Operator	Shortcut	Rule
<accrue>	,	matching documents must contain at least one of the words entered but the more unique words, the better.
<and>		matching documents must contain all of the words
<or>		matching documents must contain at least one of the words
.....		
<any>		same as <or> but without weighting
<all>		same as <and> but without weighting



Relational Operators

- ◆ **Relational Operators** are used with fields

Operator	Shortcut	Rule
<contains>		the string must be found within the field
<starts>		the field must start with this value
<ends>		the field must end with this value
<matches>		the field must contain a matching string
greater than	>	the numeric field value must be greater than this number
less than	<	the numeric field value must be less than this number
equals	=	the numeric field value must be equal to this number



Natural Language Operators

- ◆ **Natural Language Operators** provide “fuzzy searching” capabilities

Operator	Shortcut	Rule
	<FREETEXT>	locates documents with content that matches the natural meaning of the words given
	<LIKE>	locates documents using a query-by-example parser

Modifiers

- ◆ The behavior of operators can be modified or enhanced
 - The **<NOT>** modifier is used to exclude documents
oil <AND><NOT> cooking
 - The **<MANY>** modifier is used to count the density of a word or phrase topic within a document
<MANY><WORD> earning
 - The **<CASE>** modifier is used to perform a case-sensitive retrieval on a word
<CASE><WORD> NeXT
 - The **<ORDER>** modifier is used to indicate the order of the words you have entered are important
diver <ORDER><NEAR/5> kills <ORDER><NEAR/5> shark



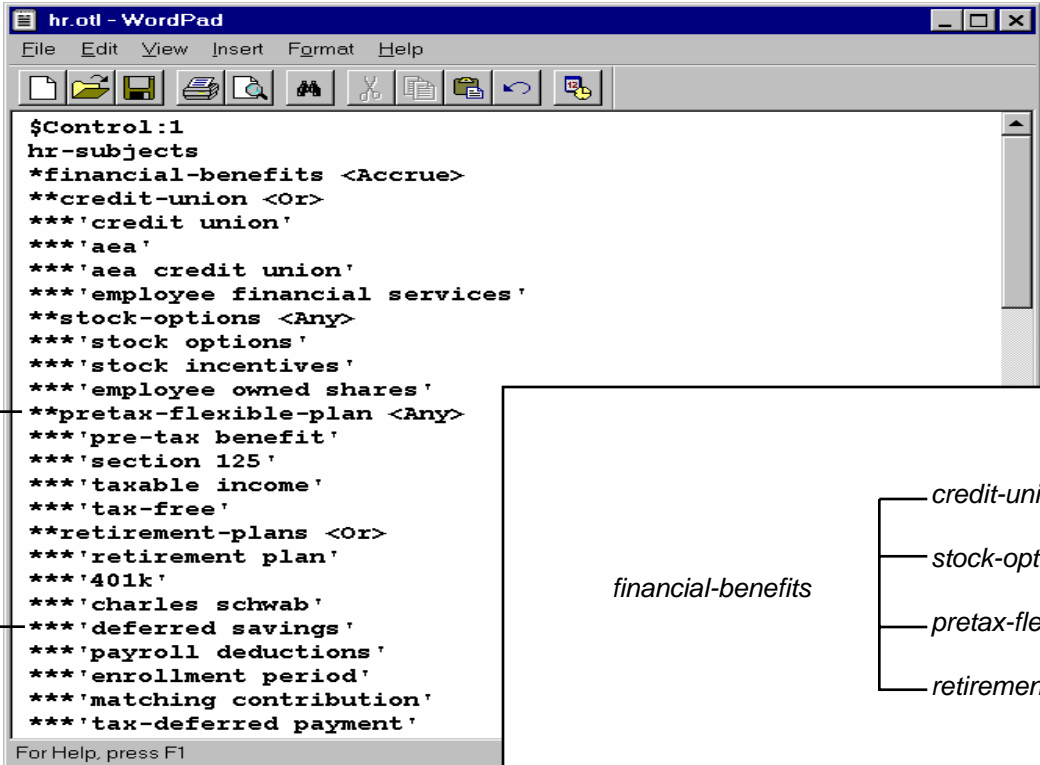
The Importance of Topics

- ◆ You can dramatically enhance the value of your applications and simplify searching for your users by incorporating topics
- ◆ Topics represent proven knowledge about a particular subject
 - The more complex a subject, the greater the value of the topic
 - Differing points of view or levels of expertise can be addressed by your topics
 - Searching is so much more effective as even novices benefit from expert knowledge captured in the topics
- ◆ Topics are cost effective, eliminating redundant work and speeding you to the right information
- ◆ They can include words, phrases, field values, and the vast array of relationships between them
- ◆ Topics include all of the components of the query language (operators, modifiers, and weights)
- ◆ Topics can be indexed against collections. Searches using indexed topics are more than 30 times as fast as those without.

Creating an OTL File

- ◆ You can create a topic outline file
 - Using any text editor
 - Using the Topic Editor

A new Java Topic Editor is under development



The screenshot shows a WordPad window titled "hr.otl - WordPad" with a menu bar (File, Edit, View, Insert, Format, Help) and a toolbar. The text content is as follows:

```
$Control:1
hr-subjects
*financial-benefits <Accrue>
**credit-union <Or>
***'credit union'
***'aea'
***'aea credit union'
***'employee financial services'
**stock-options <Any>
***'stock options'
***'stock incentives'
***'employee owned shares'
**pretax-flexible-plan <Any>
***'pre-tax benefit'
***'section 125'
***'taxable income'
***'tax-free'
**retirement-plans <Or>
***'retirement plan'
***'401k'
***'charles schwab'
***'deferred savings'
***'payroll deductions'
***'enrollment period'
***'matching contribution'
***'tax-deferred payment'
```

Annotations on the left side:

- topic in this file**: points to the line `**pretax-flexible-plan <Any>`
- phrases or words**: points to the line `***'tax-free'`

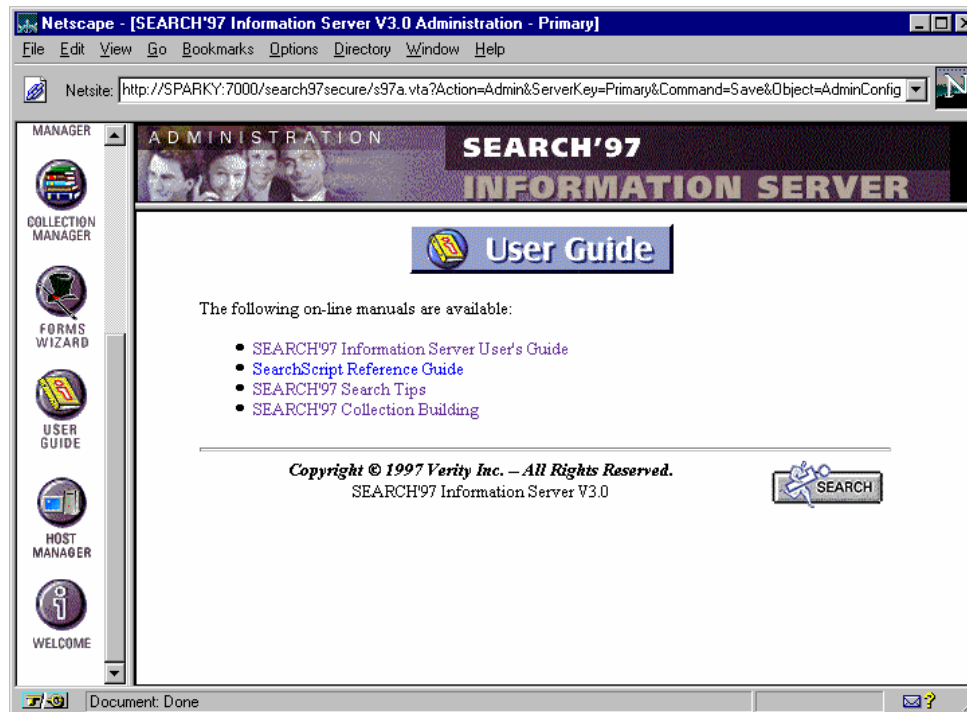
Diagram on the right side:

- A box labeled *financial-benefits* has four lines extending to the right, each pointing to a sub-topic:
 - credit-union*
 - stock-options*
 - pretax-flexible-plan*
 - retirement-plans*

For Help, press F1

Search Tips Online Guide

- ◆ This online guide provides users with advice on how to conduct organized searches for specific information
 - Operators and modifiers are introduced through a variety of search tasks which allow broadening and narrowing of searches
 - Provides practice on excluding documents





Search Form Alternatives

- ◆ It is very important to match the search form functionality to the anticipated experience level of the the user
- ◆ Add “intuition” by providing more information for those who are interested
 - Search Tips
 - Information about what is contained in document collections
 - ◆ Pre-defined queries or topics
 - ◆ Options they can set on the search form
 - How sorting works
 - What threshold is
 - Setting maximum documents
 - The difference between types of queries
 - ◆ Options they can set in terms of the results that they see
 - Setting levels of detail
 - Using advanced features of clustering and summarization



Practice Lab

- ◆ Please complete the exercises for Practice Lab #3 in your student workbook
 - These exercises will give you a chance to gain an understanding of
 - ◆ Writing queries
 - ◆ Query language basics
 - ◆ Online search training provided by Verity
 - ◆ Features of search forms