

Microsoft Exchange 2003 – Implementing a full-text Index for your Information Store

Written by Marc Grote

MCP, MCP+I, MCSA 2K/2K3, MCSA-S-E 2K/2K3, MCSE NT4/2K/2K3, MCSE-S 2K/2K3, MCT, CNA, CCNA, CCA, CCSA
mailto:grotem@it-training-grote.de

Introduction

Microsoft Exchange 2003 Server can create and manage full-text indexes for mailbox stores and public folders to provide users to search in message or public folder contents.

With the help of full-text indexing, the content in an exchange database is indexed with the result of faster content searching.

This article gives you best practices for deploying full-text indexing with Exchange 2003 Server.

This article is based on Windows 2003 Enterprise Edition (Build 3790) and Microsoft Exchange Server 2003 (Build 6944.4).

Let's begin

There are two types of search methods:

- ? Character based searches
- ? Full-text Indexing

With Character based searches, all users are searching through the e-mail messages character by character.

Full-text indexing is a powerful search tool that functions differently from other types of searches, such as character-based searches. With full-text indexing, you create an index for selected mailbox and public folder stores and then you make the index available for users to search within Outlook. Full-text indexing searches are faster than the character-based searches.

What Data is Indexed

When you deploy full-text indexing, you select the individual public folder or mailbox store to be indexed. Users can then conduct full-text searches on the messages and attachments contained in the public folder or mailbox store.

By default, the index contains the following:

- ? Subject and body of a message
- ? Names of the sender and recipient
- ? Any names that appear in the CC and BCC fields

The index also includes text from the following types of attachments:

- ? .doc
- ? .xls
- ? .ppt
- ? .html
- ? .htm
- ? .asp
- ? .txt
- ? .eml

Binary attachments, such as pictures and sounds, are not indexed but it is possible to extend the indexed file types.

Search results are only as accurate as the last time the index was updated. As the content of public folders or mailbox stores changes, the index must be updated to reflect the new content. Index updates can be performed manually or automatically on a schedule.

Setting the System resource usage

As a first step in deploying a full-text index I recommend the setting of the System resource usage.

Open the Exchange System Manager (ESM) expand your administrative group, select your server in the server container and select the System resource usage under “Full-Text Indexing”.

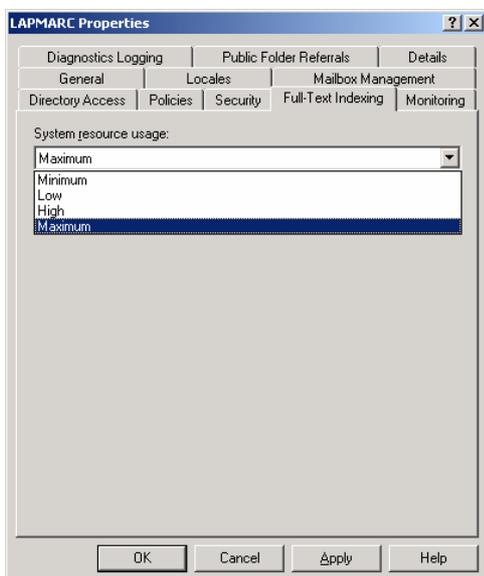


Figure 1: Select the System resource usage

Next, we have to create the first Full-Text Index

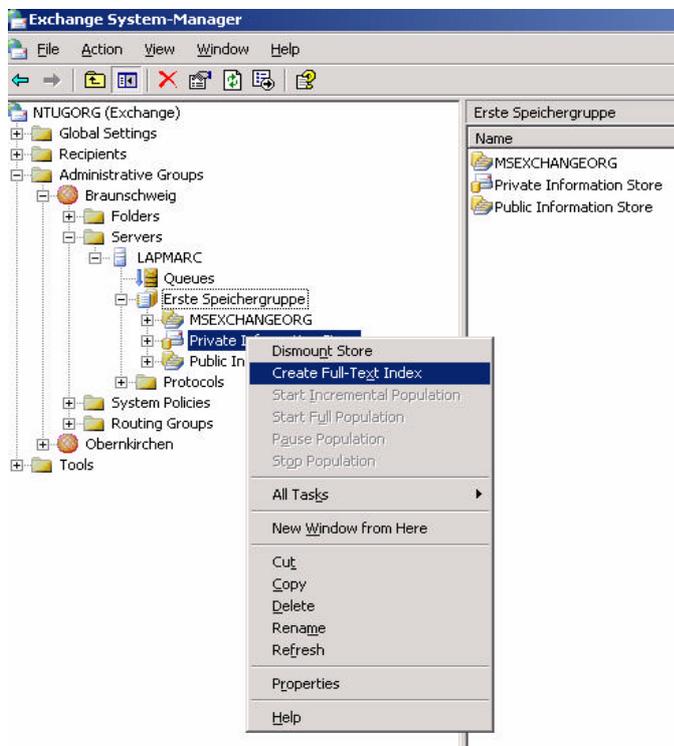


Figure 2: Create a Full-Text Index

Specify the location of the catalog



Figure 3: Set the path for the catalog

Disk Space Requirements in Exchange

The Microsoft Search service (MSSearch) powers the full-text indexing. MSSearch processes the computer running Exchange Server requests to define and populate indexes for the specified mailbox and public folder stores. MSSearch also processes queries initiated by users when they conduct full-text searches.

MSSearch requires that the disk containing the index, also called a catalog, has 15 percent free disk space at all times. Depending on the types of files you store, the size of your index can range from 10 percent to 30 percent of the size of your database.

Recommended Disk Configuration

Gather logs are the log files that contain log information for the indexing service. One set of logs exists for each index but there are even more File Types:

- ? Catalogs are the main indexes. There is only one catalog for each mailbox store or public folder store in Exchange.
- ? Property store is a database that contains various properties of items indexed in the catalog. There is only one property store per server.
- ? Property store logs are the log files associated with the property store database.
- ? Temporary files are the files that contain temporary information used by MSSearch.
- ? Gather logs are the log files that contain log information for the indexing service. One set of logs exist for each index.

Microsoft recommends using a RAID-0+1 (RAID 1+0 = RAID 10) configuration.

A RAID-0+1 disk array provides the highest performance while ensuring redundancy by combining RAID-0 and RAID-1 technologies. . RAID-0 is a striped disk array in which each disk is logically partitioned in such a way that a “stripe” runs across all the disks in the array to create a single logical partition. RAID-1 is a disk array in which two disks are mirrored. In a RAID-0+1 disk array, data is mirrored to both sets of disks, and then striped across the drives. Each physical disk is duplicated in the array. RAID 10 requires many disks but is the best RAID solution for maximum performance and redundancy. Important: RAID-5 is not recommended for full-text indexing.

Memory Requirements

Microsoft doesn't recommend running full-text indexing with less than 512 MB.

Add an additional 256 MB of RAM to the recommended configuration for a computer running Exchange 2003 Server.

Now let's go back to the configuration ...

The next step we have to specify the update interval of the index. New messages and public folder contents will be available after the update interval of the index. You have to specify an update interval for your organization. Please keep in mind that the creation and rebuilding of an index can be a processor and time intensive task.

Be sure that “This index is currently available for searching by clients” is selected, else people doesn't have the possibility to search the index.

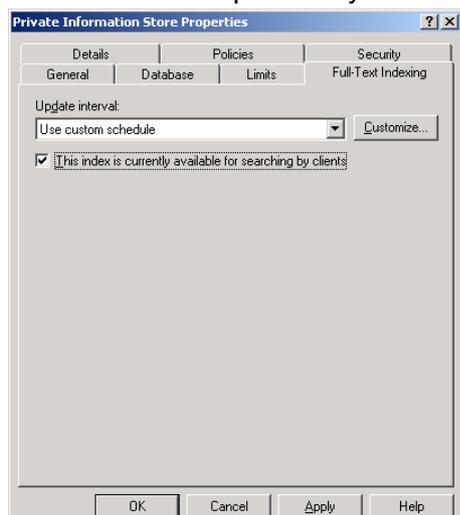


Figure 4: Select the update interval

You have the option to delete the Full-Text Index, to start the Incremental or Full Population manually. Be sure that you start a Full Population in the evening because with large and heavily accessed databases this can be a very processor and time consuming task.

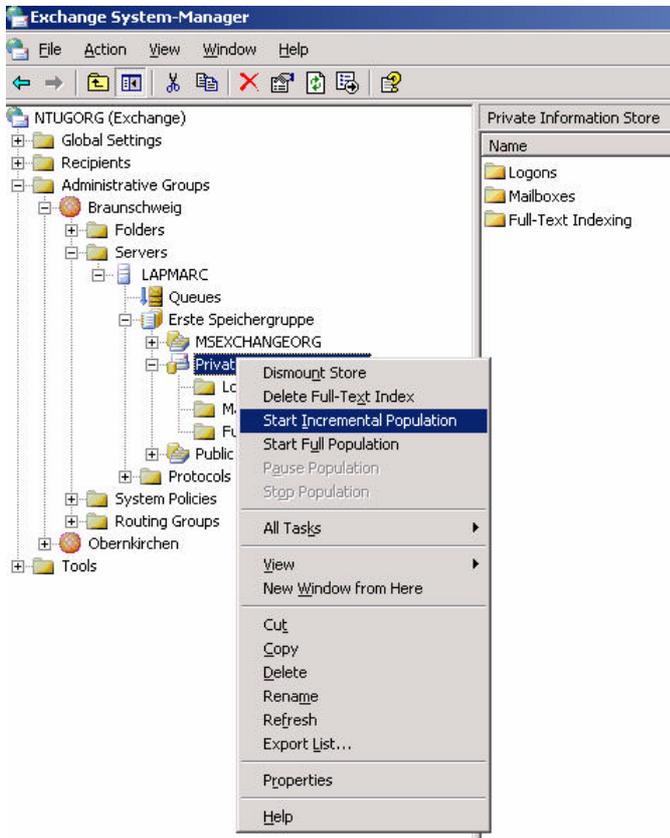


Figure 5: Maintenance Tasks

Statistics about the Full-text Index. In this picture you can see the number of indexed documents, the index state and the last build time.

An interesting point is the path to the index files.

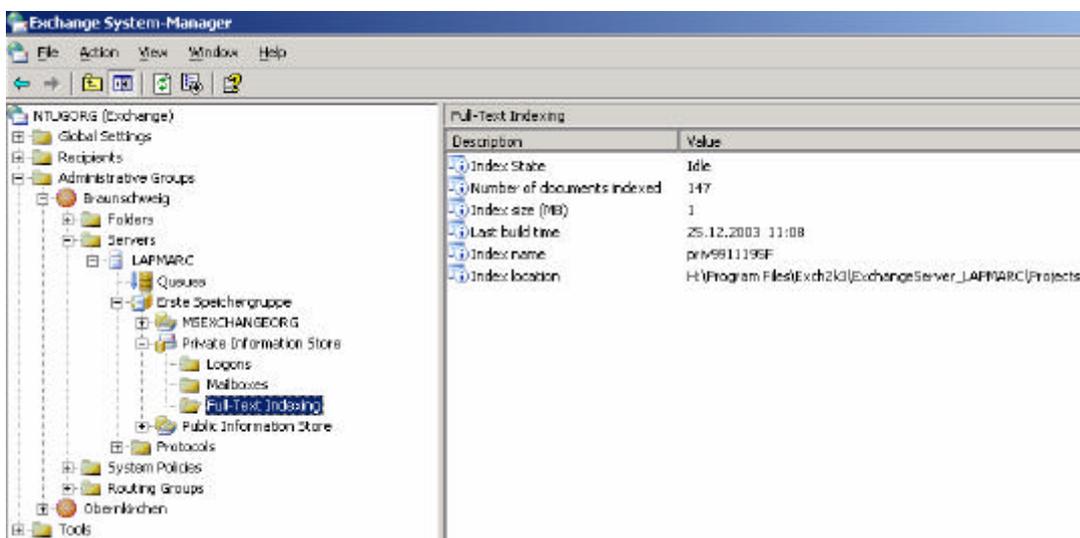


Figure 6: Statistics about the Full-text Index

How Outlook Users Access Full-Text Searching

When full-text indexing is deployed, MAPI and IMAP4 clients, such as Microsoft Outlook, can conduct full-text searches. For Outlook 2002 users, the "Advanced Find" option and the "Find option" initiates a full-text search. In particular, recently received e-mail messages do not appear in searches until they have been indexed, and full-text indexing does not search for partial word matches.

Full-Text Indexing and Multiple Languages

Exchange 2003 can create and manage full-text indexes for fast searches. Full-text indexes use the language setting of individual messages to determine which word breaker to use. Word breakers are language utilities that identify words in a document. If a message is a MAPI message, the Locale ID property is used to determine which word breaker (Dutch, English, French, German, Italian, Japanese, Korean, Spanish, Swedish, Thai, Chinese (Simplified), and Chinese (Traditional)) to use. This property value comes from the Microsoft Office Language setting.

For more information see MS KB article 325624 – "How Full-Text Indexing Works with Multiple Languages"

Noise-Word Lists

During the index and query process, Exchange indexing uses noise-word list files for each language to filter the content provided by the wordbreaker and stemmer. This noise-word list includes words and characters that MSSEARCH will **NOT** store in the catalog. This prevents MSSEARCH from storing useless information and wasting disk space. You can find the NOISE list for your language in \Program Files\exchsrvr\ExchangeServer_SERVERNAME\Config.

Modifying the Global gatherer parameter

Use the file GATHPRM.TXT to modify the global gatherer parameters. You can find the GATHPRM.TXT file in \Program Files\exchsrvr\ExchangeServer_SERVERNAME\Config.

Tools

To enhance the functionality of the full text indexing, Exchange 2003 provides some helpful tools like:

CATUTIL = Catutil.exe is an utility used to move the index files and property store, and also contains an integrity checker for the Search property store

PSTOREUTL = Can be used to change the location of the property store and the property store logs

SETTMPPATH.VBS = Sets the temp path location

GTHRLOG.VBS = Queries the gatherer status log

By default, you can find the utilities in C:\Program Files\Common Files\System\MSSearch\Bin.

Figure 7 shows the Default Location for the database files and the Gather Logs.

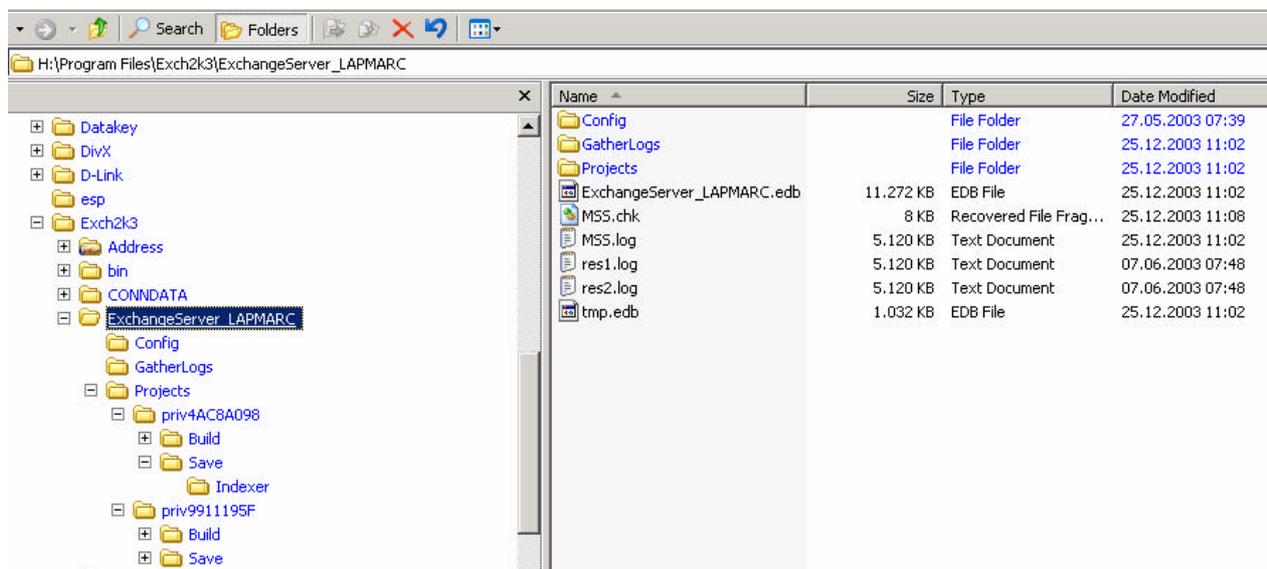


Figure 7: Default location of the Index database and the GatherLogs

Conclusion

With the help of Exchange full-text indexing, you can provide your Outlook users with a fast searching capability of their e-mails and public folders.

Related Links

Exchange 2003 Online help

Microsoft Whitepaper „Best Practices for Deploying Full-Text Indexing“ –
<http://www.microsoft.com/exchange>