DocuShare
Best Practices for Content Indexing

xerox

This document supports DocuShare Release 6.6 and greater.
Publication date: February 2012

Prepared by:
Xerox Corporation
DocuShare Business Unit
3333 Coyote Hill Road
Palo Alto, California 94304 USA

Best Practices for Content Indexing

Recommended for Sites Containing Over 500,000 Documents

The intent of this document is to provide guidance for improving DocuShare 6.6.1 search indexing performance. These instructions should be included in the planning of any new or upgraded 6.6.1 server.

Note: Refer to the DocuShare 6.6.1 *Command Line Utilities Guide* for more information on running the commands called-out in this document.

Upgrading from releases prior to 6.5

Note: Perform these procedures after upgrading to 6.6.1 and before running dsindex index_all.

Consider the needs and usage of the site, then complete **any of following commands** after upgrading to 6.6.1.

- **DSHOME\bin\countFileNumber**
  - Run this command to gather metrics on the number of documents and file sizes currently on the site. This command identifies large files that will take longer to content index.
- **DSHOME\bin\fixupMimeTypes**
  - Run this command to find files that may have been incorrectly identified with a text file MIME type.

    This command reports document and rendition objects where the MIME types assigned by the file guesser do not match the MIME type file extension. Earlier versions of DocuShare had the MIME type configuration set by default to the File Content algorithm (file type guesser) instead of the File Extension algorithm. The fixupMimeTypes command identifies documents such as *.dat files (containing many rows of numeric strings, commas, and other special characters) that may have been identified as .txt files in an earlier version of DocuShare.
- **DSHOME\bin\fixupMimeTypes -f**
  - Run this command to fix the incorrectly assigned document MIME type to match the file extension. Documents will be indexed according to the MIME type file extensions in the MIME Types table.

All DocuShare 6.6.1 servers

Note: Perform these procedures before running dsindex index_all.

Indexing documents that contain many numbers, such as spreadsheets, consumes CPU, memory, and disk space resources. All numbers and unstemmed terms not found in a standard dictionary cause additional indexing overhead. To reduce the indexing overhead caused by these types of documents and unstemmed terms, follow the recommendation below on the MIME Types table, the idoltool UnstemmedTermTree, and the AutonomyIDOLServer IndexNumbers parameters. To improve content indexing speed, use the maxfilesize configuration to restrict the document file size.

Consider the site users, their needs and search patterns, then perform **any or all** of the following recommendations:

Disk Space and Memory Allocation

- Size the server and the index resources. Know how much will be indexed.
  - Gather metrics on projected document count, projected size of content store DSHOME\documents, and estimated index size. Refer to the "**Estimating disk space**" article in the ***DocuShare Knowledge Base***.
- Configure IDOL to match available resources.
  - Set the **IDOL** configuration to match the available RAM resources on the DocuShare server. Increase performance by taking advantage of additional RAM when running IDOL on a 64-bit Windows server. Run the appropriate **idolsetup** command to configure for the available RAM.

    Note: A minimum of 6 GB of RAM is recommended for DocuShare running on 64-bit Windows servers.

  - For 4-8 GB RAM, run DSHOME\bin**idolsetup.bat..\config\idol_default.config**
  - For 8-16 GB RAM, run DSHOME\bin**idolsetup.bat..\config\idol_medium.config**
  - For 16-32 GB RAM, run DSHOME\bin**idolsetup.bat..\config\idol_large.config**
  - For 32+ GB RAM, run DSHOME\bin**idolsetup.bat..\config\idol_unlimited.config**

DocuShare Database Maintenance

- Verify the existence of database indices
  - Run **resetIndexes** to rebuild the default set of DocuShare database indices, if necessary.
- Optimize the database
  - Use either **UPDATE STATISTICS** (SQL Server), **GATHER STATISTICS** (Oracle), or **VACUUM** (PostgreSQL) to optimize the database.
  - Schedule database optimization to run daily.

Document Content Indexing

- Determine if the site requires indexing the text content of documents. Disabling content indexing will dramatically increase indexing performance and resource utilization.
    - Document metadata is always indexed. Disabling content indexing does not affect metadata indexing.
- Consider disabling content indexing if searching is targeted at metadata properties; such as Title, Subject, Keywords, Author, Description, etc.
    - Go to **Administration Menu l Services and Components l Index** to disable/enable content indexing globally for the site. Content indexing is enabled by default.

MIME Types Configuration

- Edit the **MIME Types table** at **Administration Menu l Site Management l MIME Types** to control indexing by MIME type.
    - Set **MIME Types Assignment Method** to **File Extension algorithm**. Do not use the File Content algorithm unless uploaded document filenames do not include a file type extension such as .doc, .pdf, .txt, etc.
    - Exclude from indexing document file types that do not require full text search, such as spreadsheets, binary files, images. Click **Edit** beside a MIME Type and exclude that document type from being indexed.

IDOL Server Configuration

- Adjust indexing options to balance indexing performance with the types of documents to be content indexed, and which meet the requirements of the organization and end-users. To do so, use the following steps to either run the appropriate **idoltool** command or manually edit the IDOL Server configuration file located in DSHOME\IDOLServer\IDOL**AutonomyIDOLServer.cfg**.
- **Reduce memory usage for unstemmed terms**
    - Located in the [Server] section of the AutonomyIDOLServer.cfg file, the **UnstemmedTermTree** parameter performs wildcard matching before stemming.
        - A value of **true** performs wildcard matching before stemming takes place. With true, content.exe memory usage is higher because the server is now indexing large quantities of files that contain many numbers; such as Excel files.
        - A value of **false** does not store the unstemmed terms internally for spelling correction or pre-stem wildcard/fuzzy matching. With false, content.exe memory usage is lower. This reduces the chance of getting an out-of-memory error during indexing.
    - To change the value of UnstemmedTermTree, run: **idoltool.bat -s setconfig idol Server.UnstemmedTermTree <value>**. The default value is **true**. Best results for wildcard searching is to keep the value set to **true**.

- **Set the maximum bytes of file text to be indexed per document**

  Note: The parameter **NodeTableMaxFieldLength** refers to the maximum number of bytes stored on a disk for each field in a document. A field is another name for property, such as "title", but can also include "content" (or DRECONTENT) which is the text content of a document. DocuShare indexes up to 400 KB of the extracted text content of a document. A document can be bigger than 400 KB, but it may contain only a small amount of text.

  - To change the maximum bytes of file text content that can be indexed, run: **idoltool.bat -s setconfig idol Server.NodeTableMaxFieldLength <value>**. The default value is **400000**.

    Note: Increasing this value increases indexing time.

- **Set the maximum number of terms to index per document**

  Note: The parameter **MaxIndexTermsPerDocument** refers to the maximum number of distinct terms that are indexed for any document. Set the maximum value to help reduce the effect that very long documents have on the size of the index. Once the specified maximum is reached, no new terms are indexed for that document, but further occurrences of any earlier terms will be indexed.

  - To change the maximum number of terms to index per document, run: **idoltool.bat -s setconfig idol Server.MaxIndexTermsPerDocument <value>**. The default value is **20000**.

    Note: Increasing this value increases indexing time.

- **Enable or disable the indexing of numbers**
  - In the AutonomyIDOLServer.cfg file, manually edit the **IndexNumbers** parameter that is defined under each language section.

    For example, for English:
    [english]
    IndexNumbers=1

    Valid parameter values
    0=numbers are not indexed.
    1=all numbers are indexed, whether or not they appear on their own or as part of a word. For each language, the default value is 1.
    2=numbers are indexed only when they are part of a word, such as Y2K.

- **Exclude indexing of very large files**
  - To exclude indexing the text content of files larger than a desired size, run the **setindexconfig contentIndexMaxFileSize <maxfilesize in bytes>** command. This command adds the *contentIndexMaxFileSize* parameter to the Index Server configuration file (IndexServerConfig.xml).
- **Indexing split numbers**
  - If the site has documents that contain many alphanumeric terms, such as Y2K, then the total number of indexed terms in IDOL can be reduced by setting SplitNumbers=true. This helps reduce memory and disk space usage, and increases search/index performance.

- To set SplitNumbers to true, run the command **idoltool.bat -s setconfig idol Server.SplitNumber true**.

  Note: The SplitNumbers default value is false. Setting this value to true disables wildcard searching on numbers.

Index the Site

Note: Important last step (optional if the site was indexed in 6.5.3.)

- **Run dsindex index_all**
  - After making any changes from the IDOL Server Configuration section above, run **dsindex index_all** for the changes to take effect.
  - When dsindex index_all is running while the server is in production use, the command will automatically pause to favor content indexing related to user activities. Indexing resumes when the user load is reduced.
- **Pause and restart dsindex index_all**
  - New in DocuShare 6.6.1; pause dsindex index_all with the **dsindex stop** command, then restart with **dsindex -continue index_all**.

    Note: After the initial dsindex index_all has completed, future indexing can be controlled by using the following two options; Index by classname or Batch index jobs by file size.
- **Index by classname**
  - Consider running dsindex by class type, especially if there are many documents stored as custom/cloned document objects.
  - **dsindex -classname<classname>index**
  - When indexing by classname, as a replacement for index_all, index each object class in the same order run by dsindex index_all; Users, Groups, Collections, Documents, and custom cloned objects.
- **Batch index jobs by file size**
  - Consider using the **dsindex -fsizemin<number>** and **-fsizemax<number>** options to batch index larger content files during non-peak user load times.

    Note: The contentIndexMaxFileSize <maxfilesize in bytes> configuration setting supersedes the dsindex -fsizemax value.
- **Notes for DocuShare 6.5.3 and 6.6.0 upgrades**
  - DocuShare 6.5.3 and 6.6.0 were configured with **DREREFERENCE** enabled.
  - DocuShare 6.6.1 is configured with **DREREFERENCE** disabled.
  - This change reduces memory usage and startup time of content.exe. To take advantage of this change a site upgraded from 6.5.3 or 6.6.0 must run the regeneratematchindex command. This will regenerate the indices and should be done at a convenient time when it will not affect users.
  - Back up the IDOL index data in the DSHOME\IDOLServer\IDOL\content directory before running this command.
  - Skip this step if dsindex index_all was completed after upgrading to 6.6.1.

Additional Best Practices

Do not use FORCE, KILL, or End Task

- After issuing the command to stop DocuShare, do not use **Force**, **Kill**, or **End Task** to shut down the content.exe process. Doing so will corrupt the search indices and require a re-indexing of the site to restore the index.
- Note that content.exe may not shut down with the stop_docushare command if content.exe is busy indexing. If this happens, the only option is to wait until content.exe finishes and shuts down. When DocuShare restarts, the system checks the status of content.exe. If content.exe is busy, DocuShare will not restart until content.exe has completed indexing and shuts down.

Schedule automatic backup of the IDOL index data

- As part of an overall backup/recovery plan for a DocuShare server, include idoltool backup commands to back up the IDOL search indices.
- To schedule automatic backups of the Index Data at regular intervals,edit the **AutonomyIDOLServer.cfg** and add a [Schedule] section, if not present, and set the following parameters:
  - [Schedule]
  - Backup=true
  - BackupCompression=true (creates compressed backup files)
  - BackupTime=00:00 (midnight; use 24 hour notation)
  - BackupInterval=0 (elapsed time in hours between backups; 0=everyday)
  - BackupMaintainStructure=true
  - NumberOfBackups=2 (backup cycles through each one)
  - BackupDir0=E:\DataIndex_Backup0 (backup0 directory location)
  - BackupDir1=E:\DataIndex_Backup1

    Note: The number of backups must correspond to the number of BackupDirN directories that is specified. Make sure that there is enough disk space for the backup files.
- After restoring the index from a backup, run the **reindexChangesSince** command to update the index. Refer to the Command Line Utilities Guide for more information.

Defrag/compact index data

- Compaction of the Index Data is similar to defragmentation and it saves space and enhances performance.
  - Compact the Index Data immediately by issuing a **DRECOMPACT** command from a local browser: http://localhost:9001/DRECOMPACT
  - Compact using idoltool: **>idoltool -s drecompact**
  - Compact using dsindex: **>dsindex max clean**
- To schedule compaction at regular intervals, edit the AutonomyIDOLServer.cfg and add the [Schedule] section, if not present, and set the following parameters:

- [Schedule]
- Compact=true
- CompactTime=23:00
- CompactInterval=24 (elapsed time in hours between compaction; 0=everyday)
- PreCompactionBackup=true (default; if you want a backup before compaction)
- PreCompactionBackupPath=./internalbackup (default; you can set it to another directory if space is an issue in the current partition)

Configure anti-virus

- Configure anti-virus software to scan only on Writes to the DSHOME\documents directory. Exclude this directory from virus scanning on Reads. Exclude all other DSHOME directories from virus scanning, including the IDOL index directories in DSHOME\IDOLServer\IDOL\content.

Indexing Benchmark

The following table shows indexing performance based on a virtualized Windows Server 2008 environment, with a repository size of approximately 115,000 objects, and using the default IDOL configuration settings.

Note: In a VMware environment, add more CPU cycles to improve indexing performance.

Server Configuration	Repository Size	Indexing Time (dsindex index_all)
Windows Server 2008 Enterprise Number of Processors = 2 Physical Memory = 8 GB Virtual Memory = 16 GB	Number of non-document objects = 15,000 Number of documents = 100,000	6 hours