ADMINISTRATION AND CONFIGURATION GUIDE

IMPORTANT NOTICE

This documentation is provided without a fee. The content of this documentation is provided for informational use only, is subject to change without notice, may contain technical inaccuracies or typographical errors, and should not be construed as a commitment of Lursoft IT. Lursoft IT may make improvements or changes in the product described in this documentation at any time without notice.

COPYRIGHT

CONTACT INFORMATION

Email: support@siets.net

Website: www.siets.net

TABLE OF CONTENTS

PREFACE

This preface is an introduction to the SIETS Administration and Configuration Guide. It defines the audience, describes the structure of this guide, and lists typographic conventions and abbreviations used throughout the guide.

This guide is compliant with the SIETS Enterprise Manager version 1.0.

This section contains the following topics:

Audience

This guide is intended for SIETS administrators.

Structure of This Guide

This guide has the following structure:

Section	Description
Introduction	Briefly introduces SIETS.
Getting Started	Describes how to start and exit the SIETS Enterprise Manager.
SIETS In Detail	Describes SIETS in detail from an administrative viewpoint.
Administering SIETS Server Environment	Describes how to administer SIETS server environment.
Configuring SIETS Storage	Describes how to configure the SIETS storage.
Running SIETS Commands	Describes how to run a SIETS command from SIETS Enterprise Manager.

Related Information

The SIETS package includes the following guides:

Title	Description
SIETS Installation Guide	Describes how to install SIETS.
SIETS Developer's Guide	Describes SIETS from an application developer's perspective and provides reference material for building customized applications based on SIETS.

Typographic Conventions

The following styles and conventions are used in this guide:

Convention	Description
Verdana	Represents command, function, file and directory names, system messages, and command-line commands.
Hyperlink	Represents a hyperlink. Clicking on this field takes you to the identified place.
Source code	Represents code.
Comment	Represents a comment or an example in the code.

Abbreviations

The following abbreviations are used in this guide.

Abbreviation	Description
API	Application programming interface.
FTS	Full text search.
HTTP	Hyper-text transport protocol.
XML	Extensible markup language.

1. INTRODUCTION

This guide describes the SIETS administration and configuration concepts and contains step-by-step instructions for administering and configuring SIETS servers using SIETS Enterprise Manager.

This section contains the following topics:

1.1. What is SIETS?

SIETS is a system for information storage and retrieval. The SIETS system consists of the SIETS server and application programming interface (API), and SIETS Enterprise Manager.

The SIETS server is an operational unit that performs information storing and retrieval tasks by executing a predefined set of commands.

SIETS API libraries are used for building information storage and retrieval applications that are specific and customized according to your company.

SIETS Enterprise Manager is a tool for administering a single SIETS server or a cluster of SIETS servers.

SIETS is designed with grid computing in mind. SIETS server instances can run on multiple logically connected computers forming a cluster that theoretically can store unlimited amounts of data and merge search results.

Nowadays, unstructured data amounts in companies are increasing very rapidly; the only way how to effectively retrieve such data from collections and, therefore, make the data usable, is full text search (FTS). Full text search is the main methodology implemented in SIETS server for information indexing and searching.

Full text search in SIETS is based on an optimized mathematical model, which ensures very high performance for searching poorly structured information in large amounts compared to traditional SQL systems. For this purpose, in SIETS, data are stored in an inverted index.

Subjects for full text search can be any unstructured data, for example, text collections, separate phrases or words in text documents, Web pages, Web addresses, several special markups for textual and numerical data, bookmarks of HTML or XML pages, domain names, SQL database entry key IDs, file names, and so on.

The following figure illustrates the SIETS system from a high level.

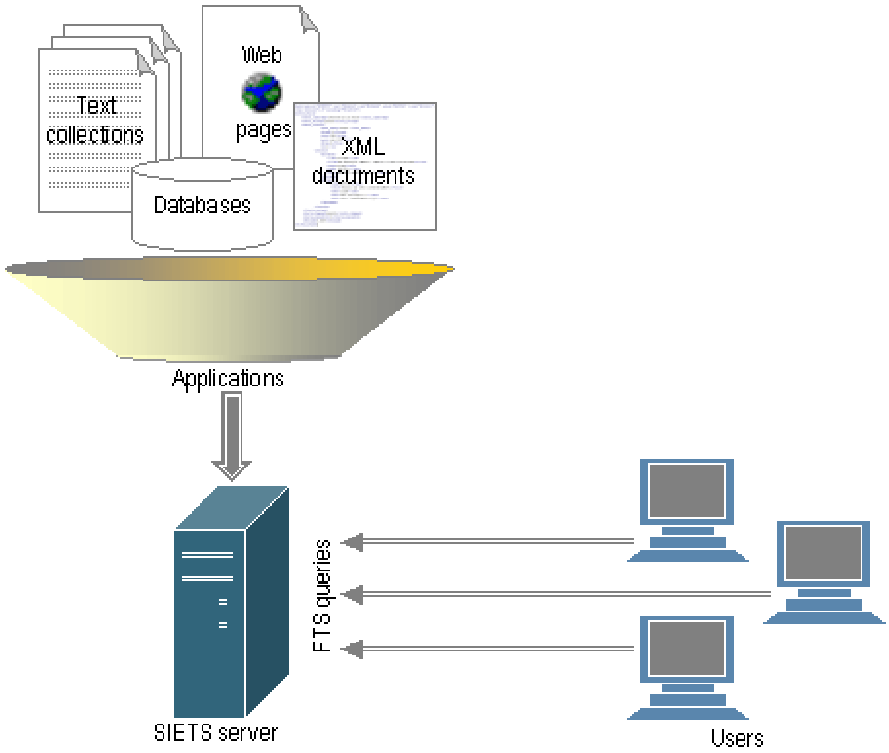


Figure 1: SIETS operational diagram

In Figure 1, users are accessing the SIETS server via FTS queries. However, also other technologies for data storing, manipulating and other implemented in SIETS, for example, retrieval queries, update requests, status and control commands, XML queries using XPath notation, and so on.

1.2. Concepts

This section contains the following topics:

- SIETS Server
- SIETS API
- SIETS Enterprise Manager
- SIETS Document
- SIETS Storage
- Vocabulary
- Document Repository
- Inverted Index

1.2.1. SIETS Server

SIETS server is a stand-alone server for storing and retrieving information such as plain texts or XML structured documents. It can be run in one or more instances per computer.

For more information, see Multiple Storages Architecture.

1.2.2. SIETS API

SIETS application programming interface (API) is a standardized set of commands for accessing the SIETS server.

1.2.3. SIETS Enterprise Manager

SIETS Enterprise Manager is an administrative tool, which allows administering and configuring all SIETS system parameters and options and which is accessed via the HTTP(S) protocol. Therefore, administering SIETS is a maximally convenient and compact task.

1.2.4. SIETS Document

SIETS document is a unit in the SIETS storage against which searching is performed. It can be unstructured or XML structured.

1.2.5. SIETS Storage

SIETS storage is a data collection for storing SIETS documents in a format that ensures a search is performed very fast. The SIETS storage is serviced by one SIETS server instance, and consists of vocabulary, document repository, and inverted index. Multiple storages can be run on a single computer.

1.2.6. Vocabulary

Vocabulary is a list of all unique words in the SIETS storage. Unique words are found in documents and added to the vocabulary while storing these documents to the SIETS storage. Each SIETS storage has its own vocabulary. Each word in the vocabulary has an ID of the integer type assigned to it. Vocabulary is stored in RAM for better performance.

1.2.7. Document Repository

Document repository is a place where all SIETS documents are stored in the format, in which they were stored in the SIETS system, for returning the documents on a search request. Each SIETS storage has its own document repository.

1.2.8. Inverted Index

Inverted index is a list of words, where each word has a list of pointers to SIETS documents in which the word occurs. Inverted index ensures fast FTS functionality with possibility to build different logical expressions when performing a search. Each SIETS storage has its own inverted index.

For more information on inverted index, see What Is Inverted Index?

1.3. Overview of Administrative Functions

The following SIETS administrator tasks can be performed:

Task	Steps described in
Administering the SIETS server or a cluster of SIETS servers, which includes the following: • managing SIETS servers that have been installed • managing SIETS storages on the SIETS servers • managing SIETS Enterprise Manager user accounts	Administering SIETS Server Environment
Configuring each SIETS storage on the SIETS servers.	Configuring SIETS Storage
Running SIETS commands to test the system.	Running SIETS Commands

2. GETTING STARTED

This section describes how to start and exit SIETS Enterprise Manager, and briefly introduces the user interface.

This section contains the following topics:

2.1. Starting SIETS Enterprise Manager

To start SIETS Enterprise Manager, proceed as follows:

1. Open the Internet browser.
2. In the **Address** field, enter the following

   http://<host>/cgi-bin/siets/sem.cgi

   where the <host> is the SIETS Enterprise Manager host name.

   If SIETS Enterprise Manager has been installed from the installation CD, in the **Address** field you can enter only http://<host>/ after which the SIETS welcome page appears. In the SIETS welcome page, click the **SIETS Enterprise Manager** link.

   The login window appears.
3. In the login window, enter the login name and password, and click **Login.**

2.2. Understanding the User Interface

When you login in the SIETS Enterprise Manager, the **Main Menu** appears.

Hyperlinks are used for navigation through different windows within SIETS Enterprise Manager.

As you navigate within SIETS Enterprise Manager, the path of where in SIETS Enterprise Manager you currently are is displayed at the top of the window.

Buttons are used for performing actions, such as submitting changes to the SIETS server, or refreshing information in SIETS Enterprise Manager.

2.3. Refreshing Information

To ensure that information displayed in SIETS Enterprise Manager is the latest from the SIETS server, in some windows the **Refresh** button is available.

2.4. Exiting SIETS Enterprise Manager

To exit the SIETS Enterprise Manager, click **Logout.**

3. SIETS IN DETAIL

This section contains the following topics:

3.1. SIETS in Corporate Networks

SIETS system can be integrated into an existing corporate network system. The SIETS server is incorporated into the network system just like any other server. The following figure describes a sample corporate network with the SIETS server.

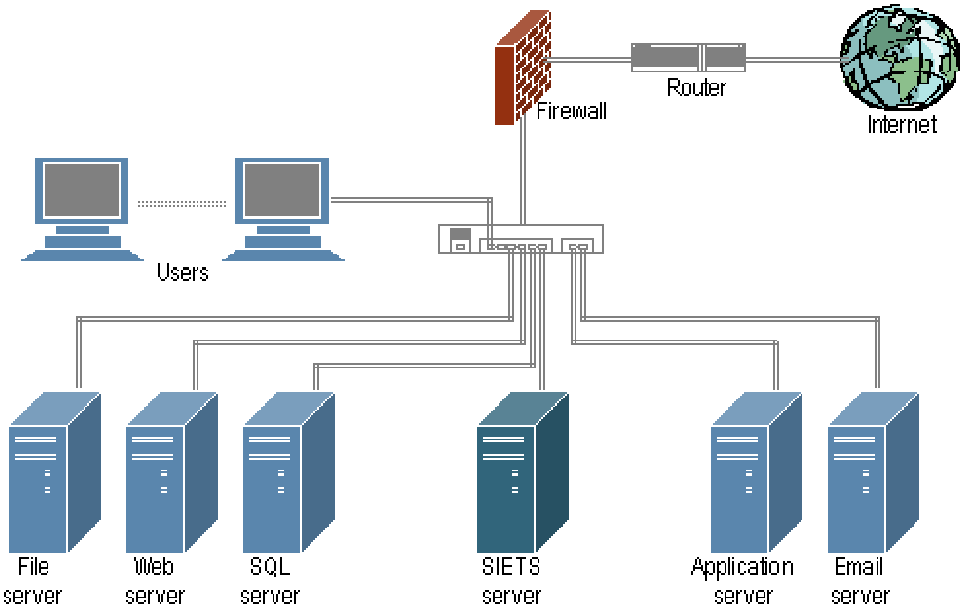


Figure 2: SIETS server in a corporate network

Application servers and transaction processors from an existing corporate network can access the SIETS server as active SIETS API clients; in that case, for security reasons, end users cannot directly access the SIETS server. In that way, SIETS server can be used in any corporate network, independently from the operation system, database environment, or programming language used for application development.

The SIETS server is software that is installed on an Unix-like operating system, for example, RedHat Linux.

In the sample presented in Figure 2, the SIETS server is installed on a single separate computer.

It is possible to install and run the SIETS server on the same computer with other services, such as Web server and application server. However, there are the following issues, if having the SIETS server installed and run on the same computer with other services:

- For large data amounts, the SIETS server intensively uses disk input and output and CPU resources, which may interrupt other services, and vice versa: other services can reduce performance of the SIETS server.
- Problem tracking becomes more complicate.

For very large data amounts, SIETS supports sever clustering, which ensures performance scalability.

For more information on SIETS server multi-server architecture, see Multi-Server Architecture.

3.2. SIETS Application Environment

As already described in the previous section, the SIETS system is used for data storage and retrieval. The SIETS system is an environment for executing data storage and retrieval commands, which are called from applications. The commands are understood and executed by the SIETS server. The applications are written by application developers. For more information on developing applications for the SIETS server, see the *SIETS Developer's Guide.*

3.3. SIETS Architecture

This section describes SIETS from various architectural perspectives.

This section contains the following topics:

- SIETS Architecture and User Roles
- Multiple Storages Architecture
- Multi-Server Architecture
- Understanding Full Text Indexing

3.3.1. SIETS Architecture and User Roles

The following figure describes SIETS architecture from user roles perspective:

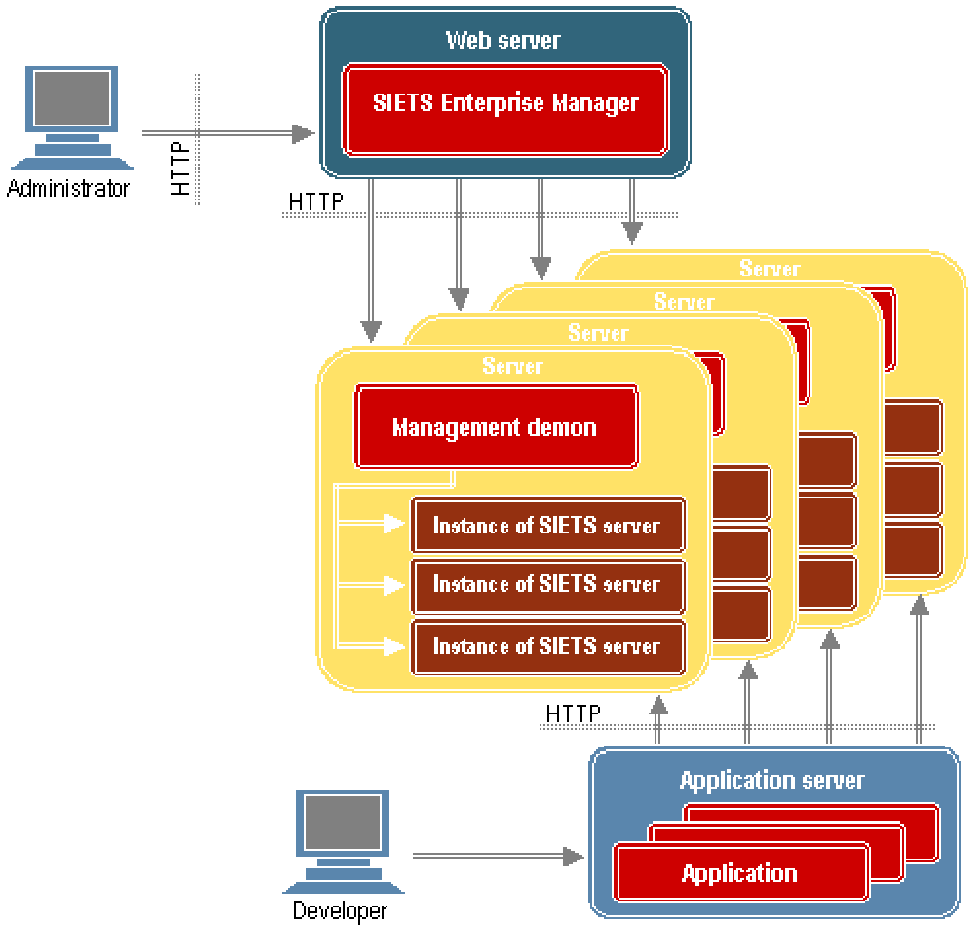


Figure 3: SIETS architecture from user roles perspective

SIETS administrator uses SIETS Enterprise Manager to administer SIETS servers. SIETS Enterprise Manager is a CGI executable installed on a Web server, which allows configuring SIETS server parameters and options. After parameters and options are configured, they are automatically submitted to the management demon of each SIETS server.

This ensures that performing SIETS administering tasks is convenient and can be done remotely.

Developers create applications, which are run on an application server, which initializes SIETS API command calls and sends them to SIETS servers via HTTP.

In Figure 3, for a better understanding of roles, the Web server and application server are on separate computers. However, the Web server and application server can be on the same computer.

3.3.2. Multiple Storages Architecture

The following figure describes how multiple SIETS server instances can be run on a single SIETS server:

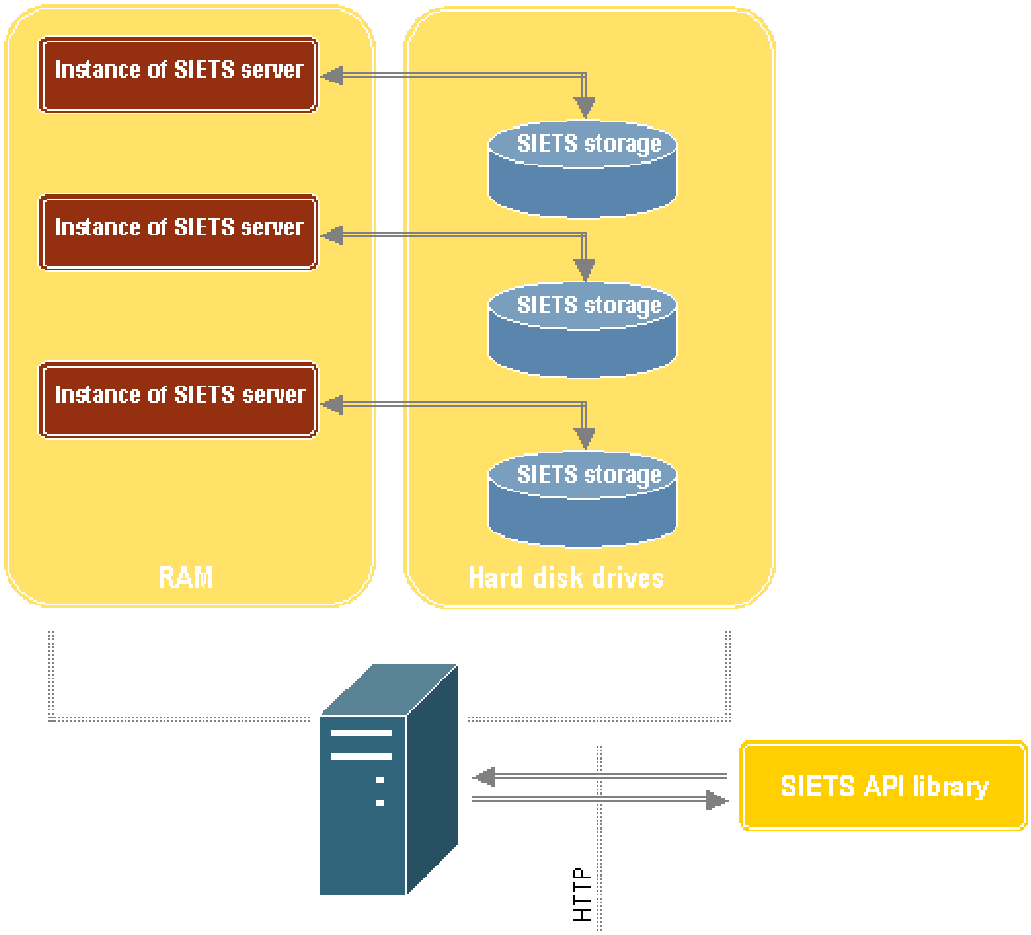


Figure 4: SIETS multiple storages architecture

Multiple instances of the SIETS server can be run on a single computer, which each works with its own SIETS storage.

3.3.3. Multi-Server Architecture

To ensure scalability of larger amounts of data, the SIETS server can be clustered sharing a single SIETS storage across many computers.

The following figure describes how a single SIETS storage can be distributed on several SIETS servers:

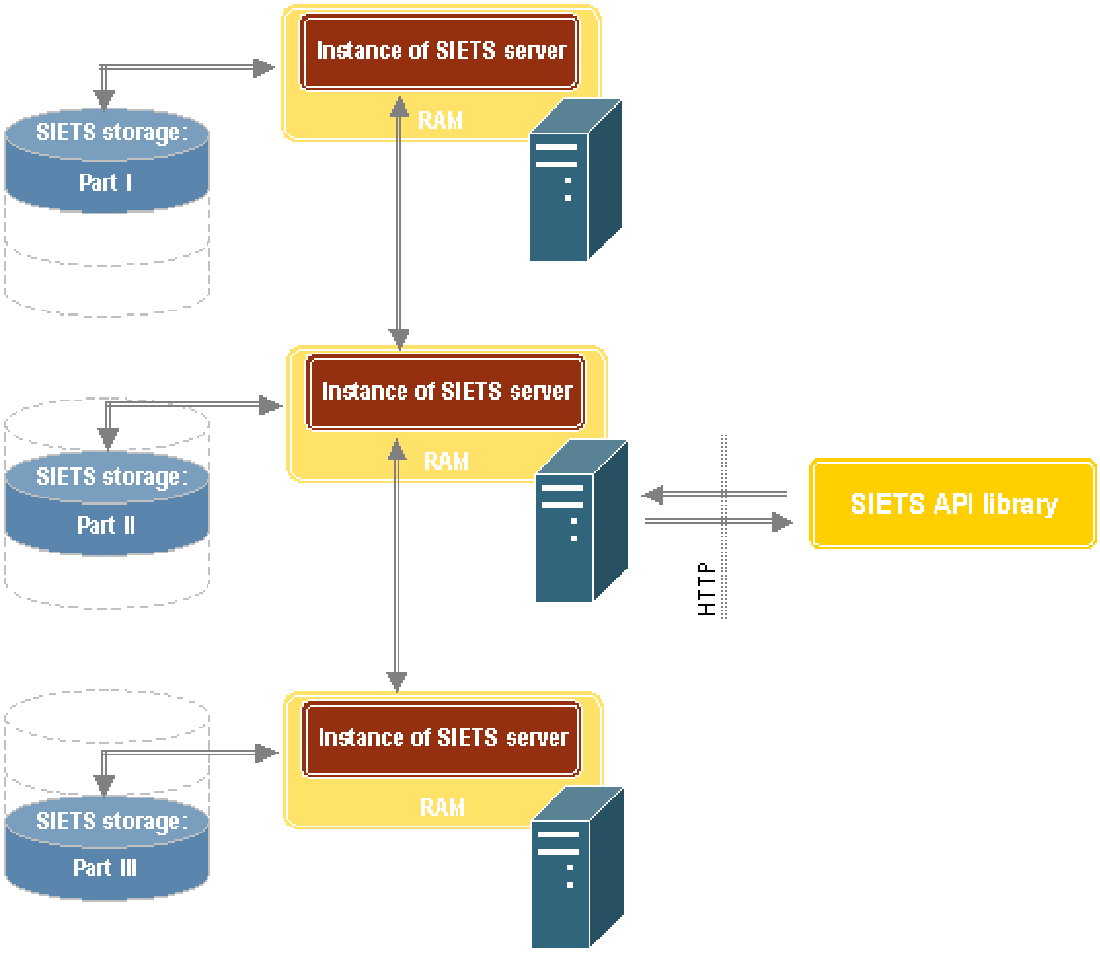


Figure 5: SIETS multi-server architecture

3.3.4. Understanding Full Text Indexing

This section contains the following topics:

- What Is Full Text Search?
- What Is Inverted Index?
- Inverted Index Construction
- Guidelines for Working with Inverted Index

3.3.4.1. What Is Full Text Search?

Full text search (FTS) is a data selection mechanism, which returns a document set containing words in a definite logical combination from a text collection.

Example:

Document Nr.	Text collection in document
1	sun, rain, cloud
2	sun, rain
3	cloud, snow, storm
4	rain, hail

Search request	Documents returned
rain	<1, 2, 4>
sun AND cloud	<1>
storm OR hail	<3, 4>
rain AND (cloud OR hail)	<1,4>
rain NOT sun	<4>

3.3.4.2. What Is Inverted Index?

Inverted index is a list of words, where each word has a list of pointers to SIETS documents in which the word occurs.

Example:

In the example from the previous section, the inverted index is as follows:

Word	Documents in which word appears
sun	1, 2
rain	1, 2, 4
cloud	1, 3
snow	4
storm	4
hail	5

What is the use of the inverted index in FTS? Let us assume that in the example we need to find documents containing the word “cloud”. To do that, we only need to find and read an entry in the inverted index, which doubtless is much faster operation than scanning the whole text collection. Now, let us assume that in the example we need to find documents containing the word “cloud” AND “snow”. To do that, we need to find entries in the inverted index and perform the set intersection operation, which also is a relatively simple and fast operation. For other logical expressions between the search terms other set operations are performed.

In reality, FTS algorithms are more complex since usually additional functionality is added, such as sorting search results by relevance according to the search query, searching for exact phrases and so on. However, the inverted index is the basis for efficient FTS.

3.3.4.3. Inverted Index Construction

A text collection can be considered as a document set, where each document consists of a set of words contained by the document, in other words, a relation <document, word>. Initially, on the disk, this set is ordered by documents. To construct the inverted index, the list must be sorted by words.

In that way, pointers to documents in which a word appears are adjacent on the disk and, therefore, can be used for searching effectively on a disk equipment using electromechanical storage technologies, for example, hard drives with mechanical disk heads.

Example:

Document sample:

| <1, sun> <1, rain> <1, cloud> | <2, sun> <2, rain> | <3, cloud> <3, snow> <3, storm> | <4, rain> <4, hail> |

Invert to the following index sample:

<1, sun> <2, sun> <1, rain> <2, rain> <4, rain> <1, cloud> <3, cloud> <3, snow> <3, storm> <4, hail>

Thus, after the document collection is loaded in the SIETS system, it is necessary to construct the inverted index. For large data amounts, this operation can be time consuming, because, although the algorithm is relatively simple, for data amounts exceeding RAM, the disk head movement is increased and can cause performance bottleneck.

In real life, rarely you have to perform FTS in a fixed data collection. This means that usually the inverted index construction is performed on a quite regular basis, which is why it should be maximally invisible to users. In SIETS the inverted index construction algorithm is designed with many rational and effective optimizations to achieve high performance, the following of which are visible and must be understood by the SIETS administrator:

- If the memory reserved for memory cache is enough for the data amount being imported, a special cache, which is located in RAM, is used for indexing. This ensures that documents are available for FTS immediately and users do not notice the indexing process. In this case the status of a storage parameter /status/matrix/pool_state = normal. The pool_state parameter can be retrieved when performing status monitoring. For more information on status monitoring, see Monitoring SIETS Storage Status and Logs.
- If the memory reserved for memory cache is not enough for the data amount being imported, a secondary level cache, which is located on the disk, is used for indexing. This means that during the indexing process some documents being indexed may not be available for FTS. In this case the status of a storage parameter is one of the following:

State	Description
/status/matrix/pool_state = expanding	Data are being added to the cache.
/status/matrix/pool_state = collapsing	Data from the cache are being committed to the inverted index.

The pool_state parameter can be retrieved when performing status monitoring. For more information on status monitoring, see Monitoring SIETS Storage Status and Logs.

3.3.4.4. Guidelines for Working with Inverted Index

According to the description of inverted index and its construction algorithm in previous sections, the following guidelines are recommended for SIETS administrator, when helping application developers to find the best strategy for performing indexing in the SIETS storage:

- Do not start any document adding, deleting, or updating command while the SIETS server is indexing the previous portion of documents and the pool_state parameter is collapsing. In other words, start document adding, deleting, or updating commands when the pool_state parameter is normal and the disk cache is not being used for indexing.
- If the disk cache is used very intensively for indexing on regular basis, consider adding more memory to the computer you have the SIETS server installed on. An exception, when an intense cache usage is considered as normal, is when you first import a large data amount to a newly created SIETS storage from an existing data source.
- Develop applications so that they do not use the disk cache intensively, to avoid slow performance.
- Indexing data can take two to four times longer period than copying such data amount on the disk. Generally, it is assumed that the number is four, and it must be considered, when developing applications. The indexing time depends on application and data amount.

4. ADMINISTERING SIETS SERVER ENVIRONMENT

Administering SIETS server environment includes configuring SIETS servers and SIETS storages, managing SIETS storages by starting and stopping running them. It also includes administering user accounts for SIETS Enterprise Manager.

Note: The SIETS Enterprise Manager user accounts must not be confused with the SIETS storage users. For more information on the SIETS storage users, see Configuring SIETS Storage.

Generally, administering SIETS servers means administering a list of hosts and IP addresses where the SIETS server is installed. This involves viewing the list, adding new SIETS server that has been installed host information to the list, editing existing SIETS server host information, or removing SIETS server host information.

The core of administering SIETS implies administering SIETS storages and its instances, which includes adding new SIETS storages, delete SIETS storages, monitoring SIETS storage status, starting and stopping instances of a SIETS storage, and running SIETS crawler tasks.

This section contains the following topics:

4.1. Administering SIETS Servers

As mentioned earlier, administering SIETS servers means administering a list of hosts and IP addresses. However, by default the local host information is already included in the SIETS servers list when installing SIETS Enterprise Manager. This implies that if your SIETS system consists of one SIETS server and one SIETS Enterprise Manager, and there are both installed on the same computer then the SIETS server host information is already added to the SIETS server list and you can skip this section and proceed to Administering SIETS Storages.

This section contains the following topics:

4.1.1. Viewing SIETS Servers List

To view a list of the SIETS servers, proceed as follows:

4. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **SIETS Servers.**

   The **SIETS Servers** window appears.

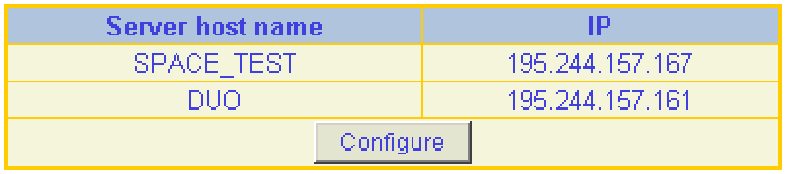


Figure 6: Managing SIETS servers

5. Review the SIETS servers list in columns described in the following table:

Title	Description
Server Host Name	Host name of the server
IP	IP address of the server.

4.1.2. Adding SIETS Servers

After you have installed the SIETS server on the hardware, you must add it to SIETS Enterprise Manager.

For information on installing the SIETS server, see the *SIETS Installation Guide.*

To add a new SIETS server, proceed as follows:

1. In the **SIETS Servers** window, select **Configure.**

   The **Server Configuration** window appears.

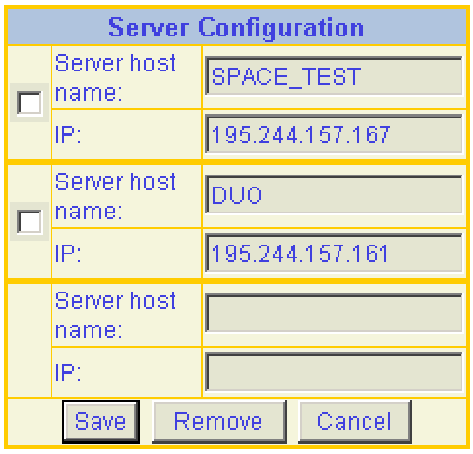


Figure 7: Adding a new SIETS server

2. In the **Server Configuration** window, at the bottom of the SIETS servers list, in the empty placeholder, enter the SIETS server parameters in the following fields:

Title	Description
Server Host Name	Host name of the server
IP	IP address of the server.

3. To finish adding the SIETS server, click **Save.**

   The **SIETS Servers** window appears with the newly added server in the SIETS servers list.

4.1.3. Editing SIETS Servers

To edit parameters of an existing SIETS server, proceed as follows:

1. In the **SIETS Servers** window, click **Configure.**

   The **Server Configuration** window appears.
2. Edit the SIETS server parameters as necessary.
3. To finish configuring the SIETS server, click **Save.**
4. To discard the changes, click **Cancel.**

4.1.4. Removing SIETS Servers

To remove a SIETS server, proceed as follows:

1. In the **SIETS Servers** window, click **Configure.**

   The **Server Configuration** window appears.
2. Select the check box on the left at the SIETS server you want to remove.
3. To finish removing the SIETS server, click **Remove.**

4.2. Administering SIETS Storages

This section contains the following topics:

4.2.1. Viewing SIETS Storages List

To view a list of the SIETS storages, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **SIETS Storages.**

   The **SIETS Storages** window appears.

Name	Status	Number of Servers	Number of Documents	Number of Words: Total/Unique	Action
web-xml	Inactive	2	0	0 / 0	Start
vocabulary	Inactive	1	0	0 / 0	Start
dvsliepaja	Inactive	1	0	0 / 0	Start
dvslursoft	Inactive	1	0	0 / 0	Start
vocabulary1	Inactive	1	0	0 / 0	Start
sabloni	Inactive	1	0	0 / 0	Start
sabloni1	Inactive	1	0	0 / 0	Start
nace	Inactive	1	0	0 / 0	Start
nace1	Inactive	1	0	0 / 0	Start
zinas	Inactive	1	0	0 / 0	Start
apollo	Processes are running partially.	1	9525	187646 / 187646	Stop
ad	Active	1	0	0 / 0	Stop
test_instance	Inactive	1	0	0 / 0	Start
siets_advert	Inactive	1	0	0 / 0	Start
apollo2	Inactive	1	0	0 / 0	Start
Refresh	Add Storage				

Figure 8: Managing SIETS storages

2. Review the SIETS storages list in columns described in the following table:

Title	Description
Name	SIETS storage name.

Title	Description
Status	SIETS storage status, which can be one of the following: Active: The SIETS storage is running. Inactive: The SIETS storage is not running. Shutting down: The SIETS storage is in the process of shutting down. Demons running partially.: Some of the SIETS storage processes are running, and some are not. For more information on running SIETS storage, see Running SIETS Storages. Instances running partially.: In case the SIETS storage is located on two or more instance, some of the SIETS servers are running, and some are not. Starting: The SIETS storage is in the process of starting and is not yet functional.
Number of Servers	Number of SIETS servers on which the SIETS storage is located. If the SIETS storage is running on a single SIETS server, then the SIETS storage has one instance running on the SIETS server. If the SIETS storage is running on more than one SIETS server, then the SIETS storage has several instances, each running on a different SIETS server. For more information on running the SIETS storage on several SIETS servers, see Multi-Server Architecture.
Number or Documents	Number of document in the SIETS storage.
Number of Words: Total/Unique	Number of words in the SIETS storage: the total number and the number of unique words.

4.2.2. Viewing SIETS Storage Instances

To view the SIETS storage instances, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage for which the instances are to be reviewed.

   The window with the selected SIETS storage instances appears.

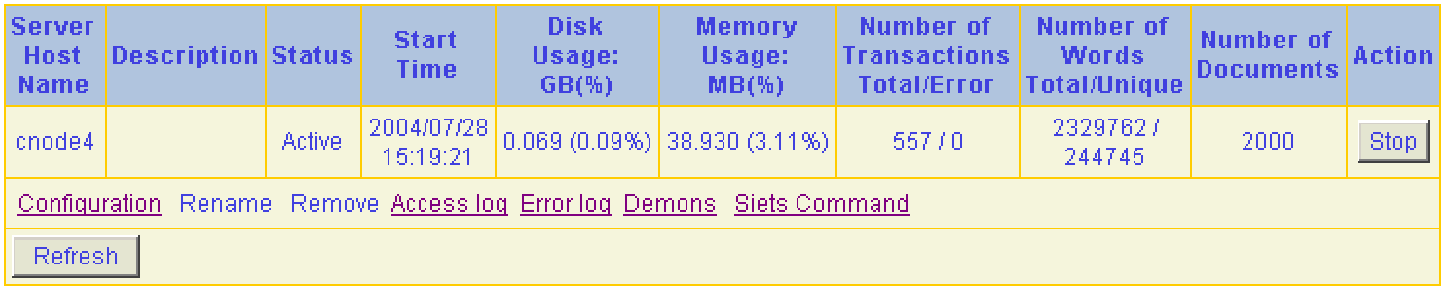


Figure 9: Viewing SIETS storage with one instance

Server Host Name	Description	Status	Start Time	Disk Usage: GB(%)	Memory Usage: MB(%)	Number of Transactions Total/Error	Number of Words Total/Unique	Number of Documents	Action
cnode4		Inactive		0.000 (0.00%)	0.000 (0.00%)	/	/		Start
Configuration Rename Remove Access log Error log Demons Siets Command									
space		Inactive		0.000 (0.00%)	0.000 (0.00%)	/	/		Start
Configuration Rename Remove Access log Error log Demons Siets Command									
Refresh									

Figure 10: Viewing SIETS storage with several instances

2. Review the SIETS storage instances in columns described in the following table:

Title	Description
Server Host Name	SIETS server on which the instance is located.
Description	SIETS storage instance description.
Status	SIETS storage status, which can be one of the following: Active: The SIETS storage is running. Inactive: The SIETS storage is not running. Shutting down: The SIETS storage is in the process of shutting down. Demons running partially.: Some of the SIETS storage processes are running, and some are not. For more information on running SIETS storage, see Running SIETS Storages. Instances running partially.: In case the SIETS storage is located on two or more instance, some of the SIETS servers are running, and some are not.
Start Time	Time when the SIETS storage instance is started, if the status is Active.
Number or Documents	Number of document in the SIETS storage.
Number of Words: Total/Unique	Number of words in the SIETS storage: the total number and the number of unique words.

4.2.3. Adding SIETS Storages

To add a new SIETS storage, proceed as follows:

1. In the **SIETS Storages** window, click **Add Storage.**

   The **Add New Storage** window appears.

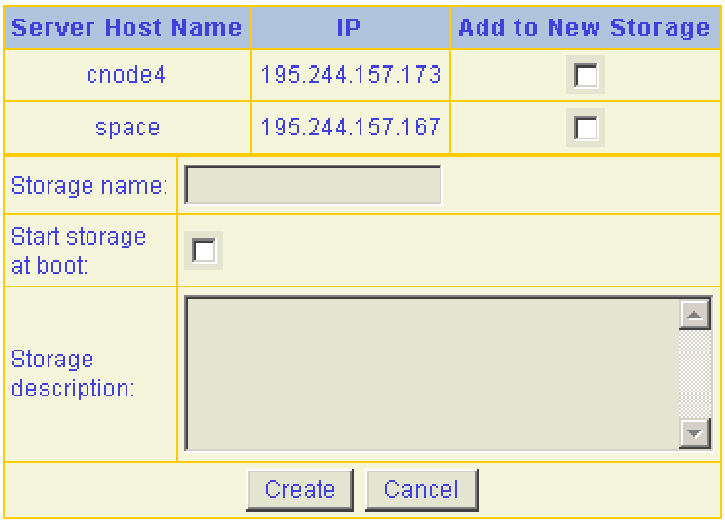


Figure 11: Adding a new SIETS storage

2. In the **Add to New Storage** column, select the SIETS servers on which the new SIETS storage is to be created.

   On each of the selected SIETS servers, an instance of the SIETS storage is created.
3. In the **Storage name** field, enter the name of the SIETS storage.
4. To start the SIETS storage automatically at every boot, select the **Start storage at boot** check box.
5. In the **Storage description** field, enter SIETS storage description.
6. To finish adding the SIETS storage, click **Create.**

   The SIETS storage is created with instances on each of the selected SIETS servers.

   The **SIETS Storage** window appears with the newly added storage in the SIETS storage list.
7. Perform the SIETS storage configuration steps, as described in Configuring SIETS Storage.

4.2.4. Configuring SIETS Storages

Configuring SIETS storages involves the following:

- Configuring Instances
- Configuring Policy Schema
- Configuring Parameters

4.2.4.1. Configuring Instances

SIETS instance configuration concepts are described in detail in Configuring SIETS Storage.

Because configuring the SIETS storage is performed on an SIETS storage instance level, as each instance can be located on different hardware and contain different

amounts of data, configuration steps described in this section are available at the SIETS instance level.

To configure a SIETS instance, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage for which is to be configured.

   The window with the selected SIETS storage instances appears.

2. Select **Configuration** below the instance.

   The following configuration options appear:

   - Instance Configuration
   - Policy Schema Configuration
   - Parameter Configuration

3. Select **Instance Configuration.**

   The instance configuration in XML appears.

4. To edit the configuration, select **Edit.**

   The **Instance Configuration** window appears.

5. In the **Instance Configuration** window, edit the configuration as necessary.

   For more information on configuration parameters, see Configuring SIETS Storage.

6. To save changes and close the window, select **Save.**
7. To close the window without saving changes, select **Cancel.**

4.2.4.2. Configuring Policy Schema

Policy schema configuration denotes structure of documents that are to be imported to the SIETS storage in terms of what each document part meant for. You must change the default policy schema configuration only if the structure of documents that are to be imported to the SIETS storage is different from the default SIETS document structure. If you use SIETS crawlers to import and index documents, then the default SIETS is used and you do not have to change the default policy schema configuration.

To configure policy schema, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage for which the policy schema is to be configured.

   The window with the selected SIETS storage instances appears.

2. Select **Configuration** below the instance.

   The following configuration options appear:

   - Instance Configuration
   - Policy Schema Configuration
   - Parameter Configuration

3. Select **Policy Schema Configuration.**

   The policy schema in XML appears.

4. To edit the configuration, select **Edit.**

The **Policy Schema Configuration** window appears.

5. In the **Policy Schema Configuration** window, edit the configuration as necessary.
6. To save changes and close the window, select **Save.**
7. To close the window without saving changes, select **Cancel.**

4.2.4.3. Configuring Parameters

To configure SIETS instance parameters, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage for which the policy schema is to be configured.

   The window with the selected SIETS storage instances appears.

2. Select **Configuration** below the instance.

   The following configuration options appear:

   - Instance Configuration
   - Policy Schema Configuration
   - Parameter Configuration

3. Select **Parameter Configuration.**

   The **Instance Parameters** window appears.

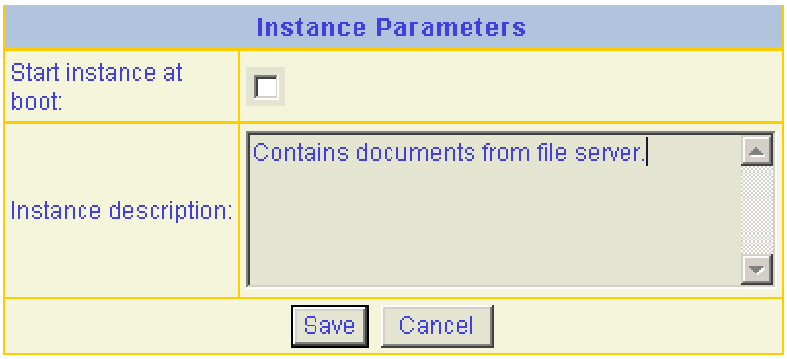


Figure 12: Configuring SIETS instance parameters

4. In the **Instance Parameters** window, enter parameters as described in the following table:

Title	Example
Start instance at boot	Information whether the SIETS instance is started at boot of the computer.
Instance description	Description of the instance.

5. To save the changes and close the window, select **Save.**
6. To close the window without saving changes, select **Cancel.**

4.2.5. Renaming SIETS Storages

Renaming the SIETS storage involves performing the rename operation for each instance of the SIETS storage. If the SIETS storage has one instance, then there are no problems with renaming it. However, if the SIETS storage has several instances, then you must be very careful: as renaming the SIETS storage is performed on the instance level, you must ensure that you enter exactly the same new storage name for each instance.

To rename the SIETS storage, proceed as follows:

1. Ensure that the SIETS storage to be renamed has the inactive status.

   Only SIETS storages that are not running, in other words, has the inactive status can be renamed.

   For more information on the SIETS storage status, see Running SIETS Storages.

2. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage to be renamed.

   The window with the selected SIETS storage instances appears.

3. To rename the SIETS storage instance, select **Rename** below the instance.

   The **Rename Instance** window appears.

Figure 13: Renaming SIETS storage instance

4. In the **New Storage Name** field, enter the new storage name.
5. To save the changes, click **OK.**
6. To discard the changes, click **Cancel.**
7. If the SIETS storage is located on more than one instance, to finish renaming the SIETS storage, repeat steps from 3 through 5 for each SIETS storage instance.

Note: For each SIETS storage instance, precisely the same name must be entered.

4.2.6. Removing SIETS Storages and Instances

There may be cases, when some of the SIETS storage instances must be removed, and there may be cases, when the whole SIETS storage must be removed. In either case, the removal process mainly involves removing SIETS storage instances one by one: whether it is only one instance, or all SIETS storage instances.

To remove the SIETS storage or a SIETS storage instance, proceed as follows:

1. Ensure that the SIETS storage to be removed has the inactive status.

   Only SIETS storages that are not running, in other words, has the inactive status can be removed.

   For more information on the SIETS storage status, see Running SIETS Storages

2. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage to be removed, or for which one of the instance is to be removed.

   The window with the selected SIETS storage instances appears.
3. To remove an instance, select **Remove** below the instance.
4. To confirm removing the instance, in the **Remove Storage** window, click **OK.**
5. To finish removing the instance, in the message that appears, confirm removal once again by clicking **OK.**
6. If the SIETS storage is located on more than one instance, to finish removing the SIETS storage, or to remove more SIETS storage instances, repeat steps from 3 through 5 for each SIETS storage instance to be removed.

4.3. Monitoring SIETS Storage Status and Logs

Similarly as with SIETS storage administering tasks described previously, also the status monitoring for the SIETS storage is performed on the SIETS storage instance level.

Monitoring the SIETS storage status, which is usually the main monitoring task, is performed by running the SIETS command `status`. For information on running SIETS commands, see Running SIETS Commands.

This section contains the following topics:

4.3.1. Viewing Access Log

Access log contains records for all requests sent to the SIETS storage.

To review the SIETS storage access log, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage, for which the access log is to be reviewed.

   The window with the selected SIETS storage instances appears.
2. To review the access log of an instance, select **Access log** below the instance.

   The **Access Log** window appears.

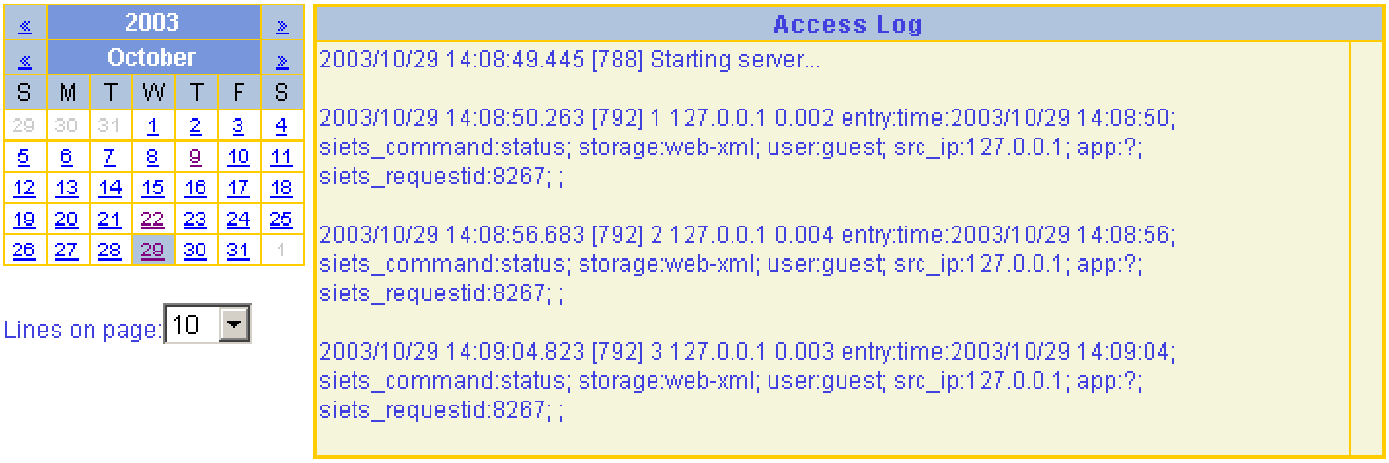


Figure 14: Viewing the access log

3. On the left, in the calendar, select the date, for which the access log is to be reviewed.

   By default, the current date is selected.

4. In the **Lines on page** drop-down list box, select the amount of log records to be displayed on a page.

   By default, ten log records are displayed on a page.

5. Review the access log records as described in the following table:

Title	Description	Example
Date	Date when the request is performed.	2003/10/29
Time	Time when the request is performed.	14:08:50.263
Reference number	Internal request reference number. Used only by the SIETS technical support.	[792]
Server request ID	Unique request ID on the SIETS server side, which is also recorded in the access log. Use this ID to search for specific records in the access log.	1
IP	IP address from which the request is submitted.	127.0.0.1
Time period	Time period on the SIETS server in seconds in which the request is executed.	0.002
SIETS command	The SIETS command being executed.	siets_command:status;
SIETS storage	The SIETS storage in which the request is performed.	storage:web-xml;
User name	Request performer user name.	user:guest;
Application request ID	Request ID on the application server side. This ID is generated by the application and returned in the SIETS reply and recorded in the access log. Use this ID to search for specific records in the access log.	siets_requested:8267;;

6. If there are more access log records than can be viewed on one page, then use the arrow on the right to navigate through the access log.

The latest records are displayed at the bottom of the access log, and by default these are displayed first, when opening the access log.

4.3.2. Viewing Error Log

Error log contains all error records for the SIETS storage.

To review the SIETS storage access log, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage, for which the error log is to be reviewed.

   The window with the selected SIETS storage instances appears.

2. To review the error log of an instance, select **Error log** below the instance.

   The **Error Log** window appears.

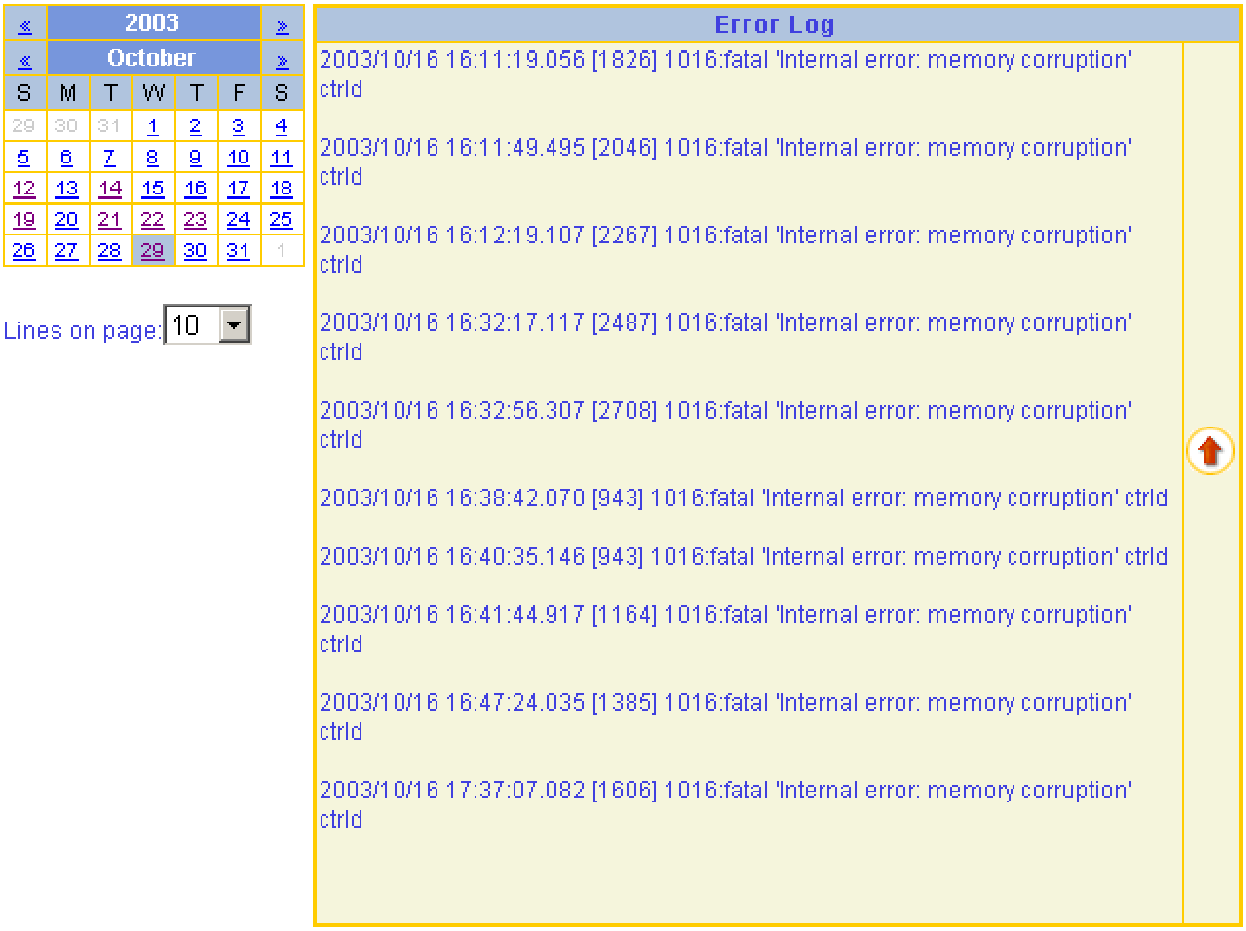


Figure 15: Viewing the error log

3. On the left, in the calendar, select the date, for which the error log is to be reviewed.

   By default, the current date is selected.

4. In the **Lines on page** drop-down list box, select the amount of log records to be displayed on a page.

By default, ten log records are displayed on a page.

5. Review the error log records as described in the following table:

Title	Description	Example
Date	Date when the request is performed.	2003/10/16
Time	Time when the request is performed.	17:37:07.082
Reference number	Internal error reference number. Used only by the SIETS technical support.	[1606]
Severity	Error severity, which can be one of the following: Warning: Returned when the command is executed successfully, but there are some problem indications Failed: Returned when incorrect input data. Error: Returned when error in the command execution. Fatal: Returned when the system is not functioning.	fatal
Message	Error message.	'Internal error: memory corruption'
Demon	Demon, in which the error occurred.	ctrld

6. If there are more error log records than can be viewed on one page, then use the arrow on the right to navigate through the error log.

   The latest records are displayed at the bottom of the error log, and by default these are displayed first, when opening the error log.

4.3.3. Viewing Demons

To review the SIETS storage access log, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage, for which the demons is to be reviewed.

   The window with the selected SIETS storage instances appears.

2. To review the demons of an instance, select **Demons** below the instance.

   The **Demons** window appears.

Demon	Status	Demon Process ID	Memory Usage: KB	Action
siets-dicd	Inactive	0	0	Start
siets-docd	Inactive	0	0	Start
siets-mtxd	Inactive	0	0	Start
siets-ctrld	Inactive	0	0	Start
siets-load	Inactive	0	0	Start

Figure 16: Viewing demons

3. Review the demons in columns described in the following table:

Title	Description
Demon	Demon name. There are the following demons: siets-dicd — SIETS dictionary demon. siets-docd — SIETS document repository demon. siets-mtxd — SIETS management demon. siets-ctrld — SIETS control demon.
Status	Demon status. A demon can be in one of the following statuses: Active — The demon is running. Inactive — The demon is not running. Shutting down — The demon is in the process of shutting down. Exited — The demon has been exited in other way than using SIETS Enterprise Manager without an error. Exited abnormally — The demon has been exited in other way than using SIETS Enterprise Manager with an error. Review the error log. For information on error log, see Viewing Error Log.
Demon Process ID	Demon process identification number assigned by the operating system.
Memory Usage: KB	Memory usage of the demon in KB.

For information on starting and stopping demons, see Running SIETS Storages.

4.4. Running SIETS Storages

Managing SIETS storages includes starting and stopping running them, which is performed by selecting the Start and Stop actions. The availability of the actions is directly related to the SIETS storage status. The following table describes how the SIETS storage status is related to action available:

Status	Action available	Description
Inactive	Start	All instances of the SIETS storage are not running. You can start one or several instances separately, or all at once by starting the whole SIETS storage.
Active	Stop	All instances of the SIETS storage are running. You can stop one or several instances separately, or all at once by stopping the whole SIETS storage.
Demons running partially.	Stop	Some of the SIETS storage instance demons are running and some are not. You can stop one or several instances separately, or all at once by stopping the whole SIETS storage. **Note:** This is not assumed as a normal situation, it can be caused by one of the following: some of the demons could been stopped manually, or a fatal error could have occurred. Review the error log. For information on error log, see Viewing Error Log.

Status	Action available	Description
Instances running partially.	No action	Some of the SIETS storage instances are running and some are not. You cannot perform an action on the SIETS storage level. You can only stop or start instances separately.

This section contains the following topics:

4.4.1. Starting and Stopping SIETS Storages

To start or to stop the SIETS storage, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **SIETS Storages.**
2. To start the SIETS storage, in the **SIETS Storages** window, in the **Action** column, click **Start** at the respective storage.

   The SIETS storage is being started, including all SIETS storage instances and their demons. The SIETS storage status is changed to Active.

   For larger SIETS storages this may take up to several minutes, as the data are being loaded in the cache.
3. To stop the SIETS storage, in the **SIETS Storages** window, in the **Action** column, click **Stop** at the respective storage.

   The SIETS storage is being stopped, including all SIETS storage instances and their demons. The SIETS storage status is changed to Inactive.

   For larger SIETS storages this may take up to several minutes.

4.4.2. Starting and Stopping Instances

To start or stop individual SIETS storage instances, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage for which the instances are be to started or stopped.

   The window with the selected SIETS storage instances appears.
2. To start the instance, in the **Action** column, click **Start** at the respective instance.

   The instance is being started, including demons of the instance. The instance status is changed to Active.
3. To stop the instance, in the **Action** column, click **Stop** at the respective instance.

   The instance is being stopped, including demons of the instance. The instance status is changed to Inactive.

4.4.3. Starting and Stopping Demons

Apart from the previous tasks described in the Running SIETS Storages section, starting and stopping demons is not considered as a common administrator task and should not be performed if not advised by the SIETS support team.

To start or stop individual demons of SIETS storage instances, proceed as follows:

1. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage for which the instance demons are be to started or stopped.

   The window with the selected SIETS storage instances appears.
2. To review the demons of an instance, select **Demons** below the instance.
3. To start a demon, in the **Demons** window, click **Start** at the respective demon.
4. To stop a demon, in the **Demons** window, click **Stop** at the respective demon.

4.5. Administering SIETS Crawler Tasks

A SIETS crawler is designed to retrieve information from the following resources:

- Web servers (HTTP)
- FTP servers
- file servers

To perform information retrieving using a SIETS crawler, SIETS crawler tasks can be created. Each SIETS crawler task can contain one or several resources from which information is to be retrieved. Also other parameters including a name of SIETS storage, on which retrieved information is stored, are set for SIETS crawler tasks.

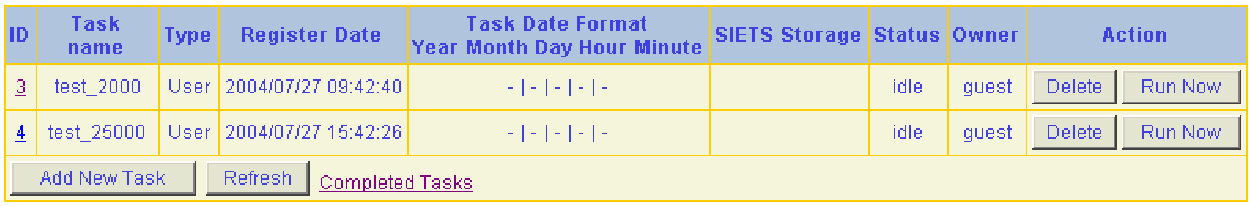
This section contains the following topics:

4.5.1. Viewing SIETS Crawler Task List

To view a list of crawler tasks, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **SIETS Crawler Management System.**

   The **SIETS Crawler Tasks** window appears.



ID	Task name	Type	Register Date	Task Date Format Year Month Day Hour Minute	SIETS Storage	Status	Owner	Action
3	test_2000	User	2004/07/27 09:42:40	- | - | - | - | -		idle	guest	Delete Run Now
4	test_25000	User	2004/07/27 15:42:26	- | - | - | - | -		idle	guest	Delete Run Now

Add New Task Refresh Completed Tasks

Figure 17: Viewing SIETS crawler tasks

2. Review the SIETS crawler tasks list in columns described in the following table:

Title	Description
ID	Task's ID, which is assigned automatically.
Task name	Task name, which can contain only alphanumeric characters.
Type	Task type, which can be one of the following: User: A user runs this task manually. Regular: The task is repeated regularly after a given time period or on given date and time. One time: The task is run once on a given date and time.
Register Date	Date and time when the task is added.
Task Date Format Year Month Day Hour Minute	If the task type is *Regular* or *One time*, then this column contains time information when or how frequent the task is run.
SIETS Storage	Storage name where results of the crawler task are stored.
Status	Current status of the task. The following statuses are possible: idle: Task is idle. crawling: Collecting data from resources is being performed. index prepare: Collected data is being prepared for indexing. indexing: Collected data is being added to full text index. canceled: The task has been stopped by selecting **Stop.** failed: The task has failed.
Owner	User's who added this task login name.
Action	Available actions for the task. The following actions are possible: Run Now: Runs the tasks. Delete: Deletes the task. Stop: Stops the task that is being run currently.

4.5.2. Adding SIETS Crawler Tasks

To add a new SIETS crawler task, proceed as follows:

1. In the **SIETS Crawler Tasks** window, select **Add Task.**

The **Create New Crawler Task** window appears.

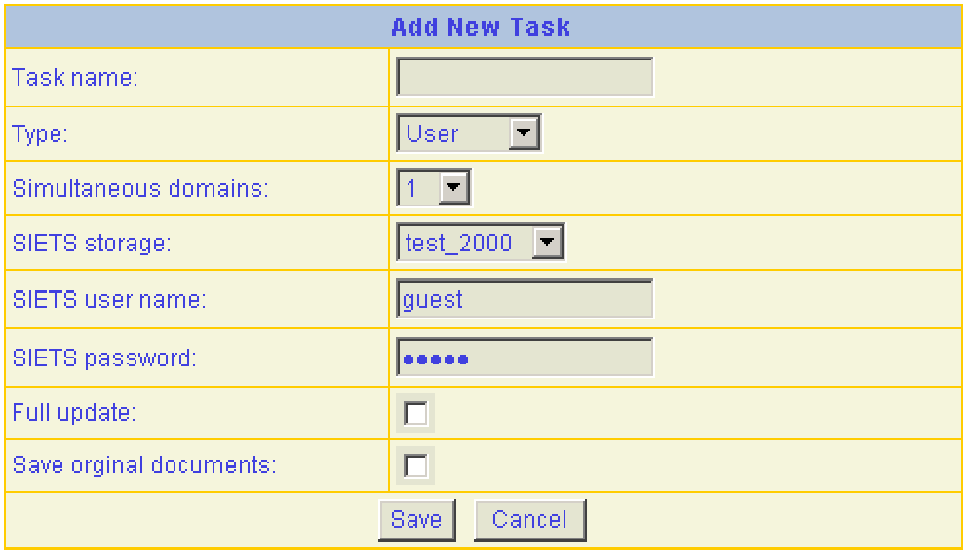


Figure 18: Adding a new SIETS crawler task

2. In the **Create New Crawler Task** window, enter the crawler task information in the following fields:

Title	Description
Task name	Task's name.
Type	Task type, which can be one of the following: User: A user runs this task manually. Regular: The task is repeated regularly after a given time period or on given date and time. One time: The task is run once on a given date and time.
Simultaneous domains	Denotes a number of domains, which are to be crawled simultaneously if several domains are to be added to this task. If you select 1 and have several domains added to the task, then all domains will be crawled sequentially. For more information on adding several domains to a SIETS crawler task, see Editing SIETS Crawler Task.
SIETS storage	Storage name where results of the crawler task are stored.
SIETS user name	User name of a user, who creates this task and is referred as to an owner of a task.
SIETS password	Password of a user, who creates this task and is referred as to an owner of a task.
Full update	Information whether the crawler task is performed on a whole domain or domains, or only changes are indexed.

Title	Description
Save original documents	Information whether original documents are stored on the SIETS storage in the base64 encoding, or is only the textual information of documents stored and links to the document returned in search result.

3. If you select *Regular* in the **Type** drop-down list box, enter the frequency information in the following fields:

Title	Description
Min	Denotes a minute of an hour or an amount of minutes of the cycle length for repeating the task. The following is possible: User: User defines a custom number of a minute of an hour in which the task is run. Values from 0 to 59 are allowed. Every 15: The task is repeated every 15th minute. Every 20: The task is repeated every 20th minute. Every 30: The task is repeated every 30th minute. **Note:** This parameter is combined with other parameters of the cycle length.
Hour	Denotes an hour of a day or an amount of hours of the cycle length for repeating the task. The following is possible: User: User defines a custom number of an hour of a day in which the task is run. Values from 0 to 23 are allowed. Every: The task is repeated every hour. Every 2: The task is repeated every 2nd hour. Every 3: The task is repeated every 3rd hour. Every 4: The task is repeated every 4th hour. Every 6: The task is repeated every 6th hour. Every 8: The task is repeated every 8th hour. Every 12: The task is repeated every 12th hour. **Note:** This parameter is combined with other parameters of the cycle length.
Day	Denotes a day of a month, a day of a week, or an amount of days of the cycle length for repeating the task. The following is possible: Month day: User defines a custom number of a day of a month in which the task is run. Values from 1 to 31 are allowed. Week day: User defines a custom number of a day of a week in which the task is run. Values from 1 to 7 are allowed. Every: The task is repeated every day. **Note:** This parameter is combined with other parameters of the cycle length.

Title	Description
Month	Denotes a month of a year or an amount of months of the cycle length for repeating the task. The following is possible: User: User defines a custom number of a month of a year in which the task is run. Values from 1 to 12 are allowed. Every: The task is repeated every month. **Note:** This parameter is combined with other parameters of the cycle length.

4. If you select *One time* in the **Type** drop-down list box, enter a precise time of when the task will be run in the following fields:

Title	Description
Min	Minute information of when the task will be run.
Hour	Hour information of when the task will be run.
Day	Day information of when the task will be run.
Month	Month information of when the task will be run.
Year	Year information of when the task will be run.

5. To finish adding the SIETS crawler task, click **Save.**

   The **Task Details** window appears.

   For more information on task details, see Editing SIETS Crawler Task.

4.5.3. Editing SIETS Crawler Task

To edit SIETS crawler task details, proceed as follows:

1. In the **SIETS Crawler Tasks** window, in the **ID** column, click ID of the task you want to edit.

   If you just added the task as described in Adding SIETS Crawler Tasks, then you do not have to perform this step.

   The **Task Details** window appears.

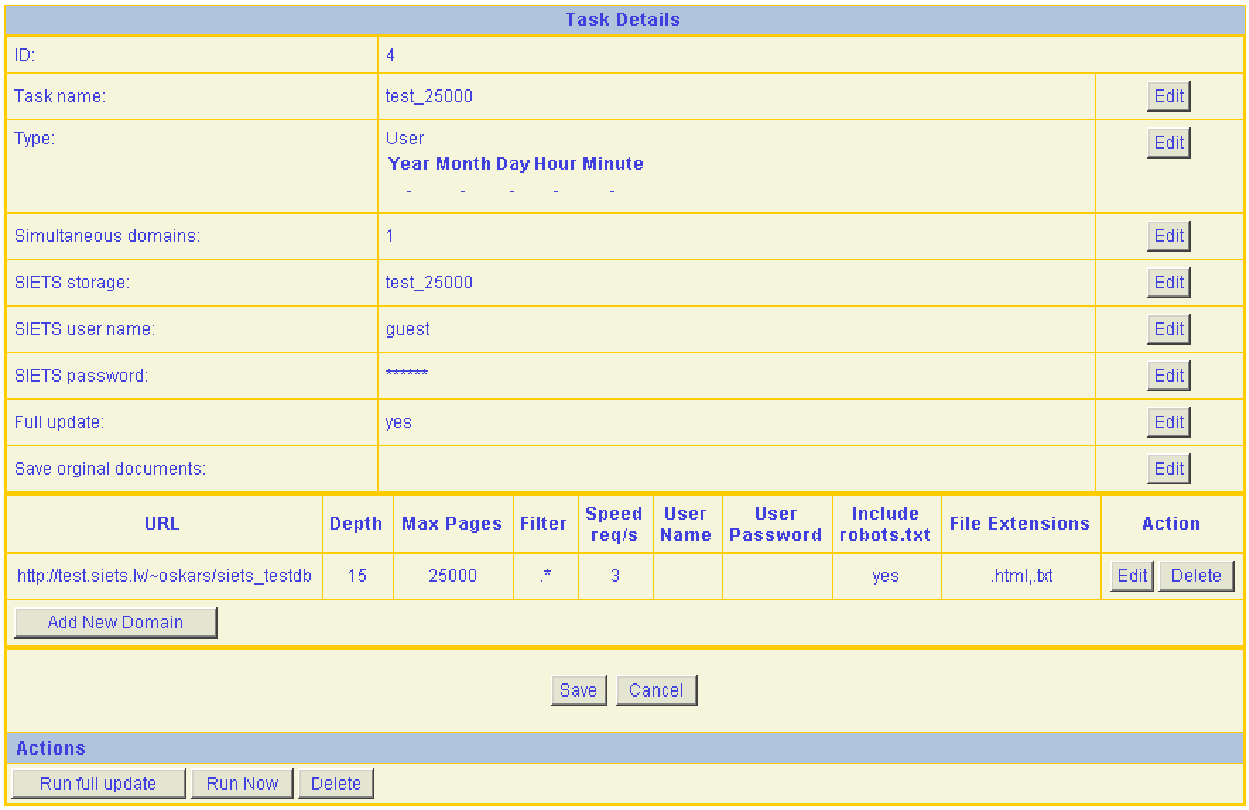


Figure 19: Editing SIETS crawler task

2. In the **Task Details** window, to edit the SIETS crawler task's parameter, select **Edit** for the parameter that you want to edit.

   The editing window appears for the selected parameter.

3. Edit task parameters as described in Adding SIETS Crawler Tasks.
4. In the editing window, to save the changes, select **OK.** To close the window without saving the changes, select **Cancel.**
5. To add a new domain, in the **Task Details** window, select **Add New Domain.**

   The **Add Domain** window appears.

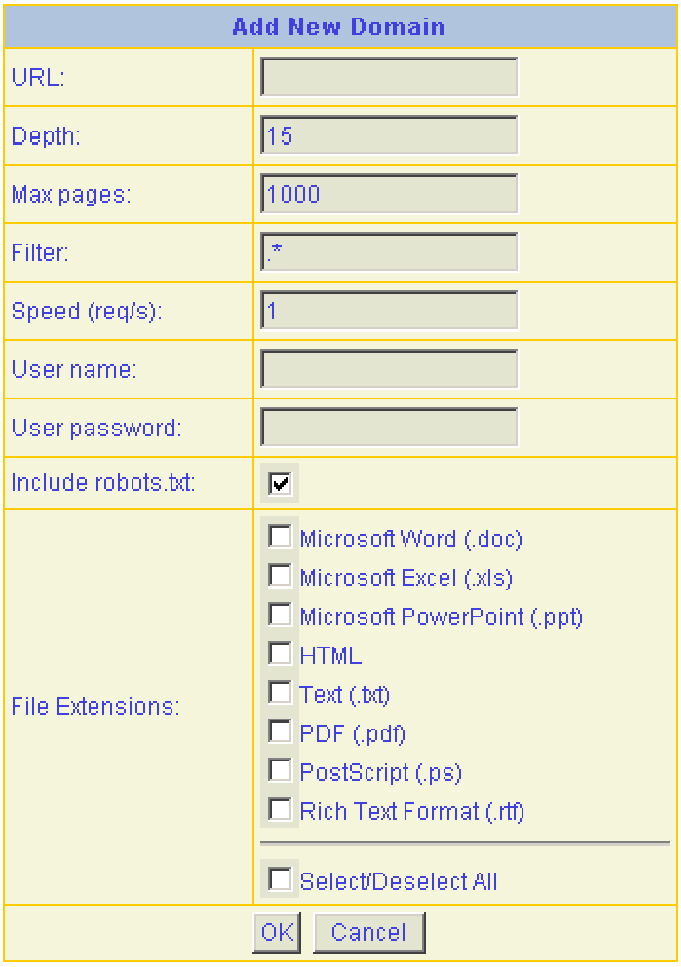


Figure 20: Adding domain

6. Enter the domain information in the following fields:

Title	Description
URL	URL of the domain in the form of <protocol>://<domain>. Protocols: http, https, ftp, sftp and file are allowed.
Depth	Depth of the domain structure, which is the maximum that the crawler is to index. If 1 is entered, the crawler will index only top level pages, if 2 is entered, the crawler will index top level pages and pages that are one level deeper than top level pages. The greater the depth number, the longer it will take the crawler to index the domain.
Max Pages	Maximum amount of pages form the domain that the crawler will index.
Filter	Regular expression that defines what pages from the domain will be indexed.

Title	Description
Speed (req/s)	Crawler's frequency of sending requests to the domain in the form of requests per second. Consult the domain administrator for what is the allowed speed for sending requests to the domain.
User name	If the domain is protected with a user name and password, the user name, which the crawler will use to connect to the domain.
Password	If the domain is protected with a user name and password, the password, which the crawler will use to connect to the domain.
Include robots.txt	Information whether the crawler is to index contents of pages and directories that are included in the domain's robots.txt file. If the check box is selected, the crawler will *include* the robots.txt file and, thus, will not index contents of the pages and directories listed in the robots.txt file.
File extensions	The crawler will index only files with extensions selected in this group.

7. To save the changes, in the **Add Domain** window, select **OK.** To close the window without saving changes, select **Cancel.**
8. The added domain appears in the domain list below the task's parameters.
9. To add more domains, repeat steps from 5 to 7 for each domain you want to add.
10. To edit a domain, select **Edit** at the domain you want to edit, and proceed as described in steps 6 and 7.
11. To delete a domain, select **Delete** at the domain you want to delete.
12. When all necessary changes for the SIETS crawler task are made, to save them, in the **Task Details** window, select **Save.**
13. To close to the **Task Details** window without saving the changes, select **Cancel.**

4.5.4. Running SIETS Crawler Tasks

A SIETS crawler tasks can be configured so that they are:

- run manually,
- repeated automatically with a given frequency, or
- run once on a given time.

For more information on setting task running frequency or a precise time, see Adding SIETS Crawler Tasks.

Also SIETS crawler tasks can be configured so that they perform full indexing of domains or only changes are indexed.

Despite the configuration settings just described in this section, there is functionality for running crawler tasks manually. Also, when running tasks manually, they can be run considering the configured full update parameter information, or they can be run on a whole domain or domains ignoring the full update parameter set.

To run a SIETS crawler task, proceed as follows:

1. In the **SIETS Crawler Tasks** window, in the **Action** column, select **Run Now** for the task you want to run.

   The task will be run considering the full update parameter set.

To run a SIETS crawler task on a whole domain or domains, proceed as follows:

1. In the **SIETS Crawler Tasks** window, in the **ID** column, click ID of the task you want to run.

   The **Task Details** window appears.

2. In the **Task Details** window, select **Run full update.**

Note: You can also run a SIETS crawler task considering the full update parameter from the **Task Details** window by selecting **Run Now.**

4.5.5. Stopping SIETS Crawler Tasks

To run a SIETS crawler task, proceed as follows:

1. In the **SIETS Crawler Tasks** window, in the **Action** column, select **Stop** for the task you want to stop.

   The confirmation message appears.

2. To confirm stopping the SIETS crawler task, select **OK.**

4.5.6. Deleting SIETS Crawler Tasks

To delete a SIETS crawler task, proceed as follows:

1. In the **SIETS Crawler Tasks** window, in the **Action** column, select **Delete** for the task you want to delete.

4.5.7. Viewing Completed SIETS Crawler Task List

A completed task list allows viewing information about SIETS crawler tasks for a selected month.

To view a completed SIETS crawler task list, proceed as follows:

1. In the **SIETS Crawler Tasks** window, select **Completed Tasks.**

   The **Completed Tasks** window appears.

2. Select the month and year of completed tasks and review the completed tasks information summary information.
3. To review a detailed information for each completed task, in the # column, select the task number.

   The **Task Details** window appears.

4.6. Administering SIETS Enterprise Manager User Accounts

The SIETS Enterprise Manager user accounts must not be confused with the SIETS storage users. For more information on the SIETS storage users, see Configuring SIETS Storage.

This section contains the following topics:

4.6.1. Viewing SIETS Enterprise Manager Users List

To view a list of the SIETS servers, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **User Accounts.**
2. In the SIETS servers list, select a server, for which the user accounts are reviewed.

   The **Users** window appears.

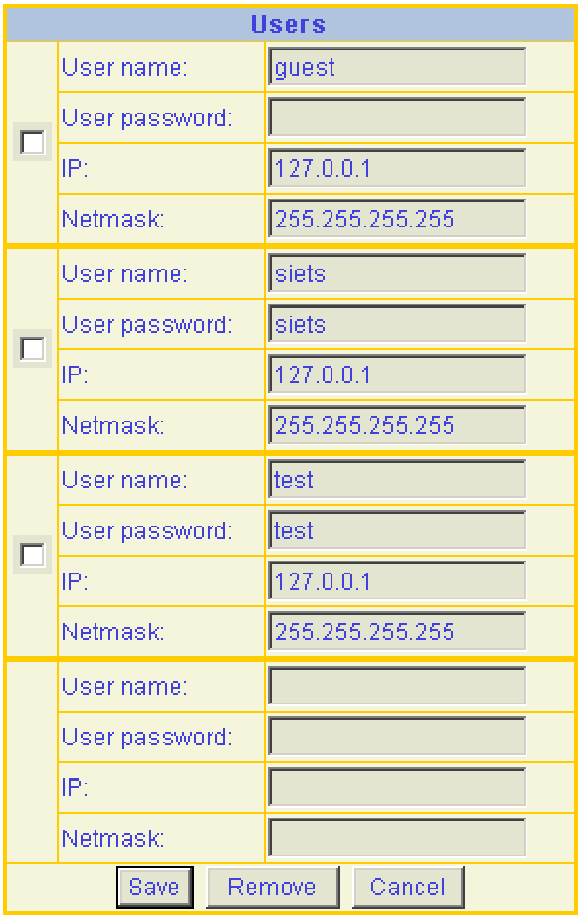


Figure 21: Viewing users

3. Review the user information in fields described in the following table:

Title	Description
User name	User login name.
User password	User password.

Title	Description
IP	User IP address.
Netmask	User netmask.

4.6.2. Adding SIETS Enterprise Manager Users

To add a new SIETS Enterprise Manager user, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **User Accounts.**
2. In the SIETS servers list, select a server, for which the user accounts are added.

   The **Users** window appears.
3. In the **Users** window, at the bottom of the users list, in the empty placeholder, enter the user parameters in the following fields:

Title	Description
User name	User login name.
User password	User password.
IP	User IP address.
Netmask	User netmask.

4. To finish adding the user, click **Save.**

4.6.3. Editing SIETS Enterprise Manager Users

To edit a SIETS Enterprise Manager user, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **User Accounts.**
2. In the SIETS servers list, select a server, for which the user accounts are edited.

   The **Users** window appears.
3. In the **Users** window, edit the user parameters as necessary.
4. To save the changes, click **Save.**

4.6.4. Removing SIETS Enterprise Manager Users

To remove a SIETS Enterprise Manager user, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **User Accounts.**
2. In the SIETS servers list, select a server, for which the user accounts are removed.

   The **Users** window appears.
3. In the **Users** window, select the check box on the left at the user you want to remove.
4. To finish removing the user, click **Remove.**

5. CONFIGURING SIETS STORAGE

When the SIETS storage is added, the SIETS storage configuration XML file is created automatically with the default SIETS storage configuration parameter values set for each SIETS storage instance.

Configuring the SIETS storage is performed on an SIETS storage instance level, as each instance can be located on different hardware and contain different amounts of data.

The SIETS storage configuration file includes parameters for SIETS storage users, and indexing options, and other parameters for performance tuning.

While you may want to leave the performance tuning parameter default values, as by default they are optimally adjusted for the performance of most common size and type of data and could be more complicate to understand, there are parameters that are specific only to your system and must be configured yourself. These are users, user passwords, optional dictionary settings, and so on.

If you, however, feel that a performance tuning is necessary, but you are not sure about which parameters must be changed, contact and get assistance from the SIETS support team.

For information on contacting the SIETS support team, see Getting Help.

SIETS storage configuration is performed in textual mode. After you select the SIETS storage configuration control, the SIETS storage configuration XML file is loaded in SIETS Enterprise Manager. Configuring SIETS storage means editing configuration parameters in the tags of the XML file.

If any of the configuration parameter tags are deleted from the SIETS storage configuration XML file, the default value of the parameter is used when running the SIETS storage.

The changes made to the SIETS storage configuration file become effective after the instance is stopped, if running, and started again with the new configuration.

For more information on stopping and starting SIETS storage instances, see Running SIETS Storages.

As opening, editing, saving, closing the SIETS storage configuration file is common for editing all configuration parameters, it is described in a separate section. Each parameter or parameter group is described in a separate section, which includes its representation in the SIETS storage configuration file and describes what it means.

This section contains the following topics:

5.1. Working with SIETS Storage Configuration File

To open, edit, save, and close the SIETS storage configuration file from SIETS Enterprise Manager, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **SIETS Storages.**
2. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage to be configured.

3. To configure an instance, select **Configuration** below the instance.
4. Select **Instance Configuration.**
5. To open the configuration file for editing, click **Edit.**
6. Edit the configuration parameter values in tags as described in the following sections.
7. To save the changes made, click **Save.**
8. To discard the changes and close the window, click **Cancel.**

5.2. SIETS Storage Configuration File Parameters

This section described SIETS storage configuration file parameters in a separate section. Each section contains the parameter and its child parameter description and example of how it they appear in the SIETS storage configuration file.

This section contains the following topics:

5.2.1. General

The following table describes general SIETS storage configuration parameters:

Second level element	Third level element	Description	Default
<general>		General information about the SIETS storage.	
	<storage>	SIETS storage name, entered when adding a new SIETS storage.	
	<port>	SIETS storage port, entered when adding a new SIETS storage.	
	<max_resultset>	Maximum number of documents returned to the result set.	1000
	<timeout>	Function timeout period in seconds. If the command is not executed during this predefined timeout period, the command returns the error.	60
	<log_path>	Relative path according to the SIETS storage directory where all log files are stored.	
	<log_rotate>	Number of days after which all log files are deleted to ensure that the disk is not over flooded with log information.	60

<table>
<tr><th>Second level element</th><th>Third level element</th><th>Description</th><th>Default</th></tr>
<tr><td></td><td><dump></td><td>Information whether the dump is to be created. The following values are possible:
<table>
<tr><td>no</td><td>No dump.</td></tr>
<tr><td>error</td><td>Only commands that cause errors are dumped.</td></tr>
<tr><td>all</td><td>All commands are dumped.</td></tr>
</table></td><td>no</td></tr>
</table>

Example:

```
<general>
    <storage>Newspapers</storage>
    <port>90</port>
    <max_resultset>1000</max_resultset>
    <timeout>60</timeout>
    <log_path>./logs</log_path>
    <log_rotate>60</log_rotate>
    <dump>no</dump>
</general>
```

5.2.2. Users

The following table describes user management SIETS storage configuration parameters:

Second level element	Third level element	Forth level element	Description
<users>			This element contains a list of SIETS storage users.
	<user>		This element contains a user name and password. It is repeated for each user.
		<name>	User name.
		<pass>	User password.

Example:

```
<users>
    <user>
        <name>John</name>
        <pass>unbreakable_password</pass>
    </user>
</users>
```

SIETS storage users must not be confused with SIETS Enterprise Manager user accounts. For information on the SIETS Enterprise Managers users, see Administering SIETS Enterprise Manager User Accounts.

5.2.3. Dictionary

The dictionary element, which is a second level element, contains several parameters and parameter groups for configuring search query defining options. This section contains the following topics, which each describes one of the search defining options:

- Special Symbols
- Wildcard Patterns

5.2.3.1. Special Symbols

The following table describes parameter for configuring special symbols:

Third level element	Description	Default
<specsymbols>	Letters and numbers are regular symbols that form words. By default, all other symbols like: ! _ % & *., are considered as word separating symbols. This element contains special symbols that are additional to the regular symbols list. The default value _ means that, for example, sleep_walk is treated as one word. Several special symbols are entered without a space or any other separator.	_

5.2.3.2. Wildcard Patterns

The following table describes parameters for configuring wildcard patterns support:

Third level element	Forth level element	Description	Default
<wildcards>		This element contains parameters for configuring wildcard patterns support.	
	<allow>	Information whether the wildcard patterns search is enabled.	yes
	<cover_ factor>	When wildcard patterns are used to define a class of words to be searched, only a limited number of statistically frequent words are searched for to ensure a higher performance. This element defines the limit in percent from the sum of all words created from the wildcard pattern appearance in the SIETS storage. **Example:** Search query: ca? All words: "car", "cat", "cap", "can", and "cab" Number of times each word appears in the SIETS storage: 4: car car car car 3: cat cat cat 2: cap cap 1: can Cover factor 60% means that shadowed words are searched and returned. 10 20 30 40 50 60 70 80 90 100 car car car car cat cat cat cap cap can Note that the word "cat" is searched, as it is important that at least one of the all appearances of "cat" fall in 60%.	95

Third level element	Forth level element	Description	Default
	<min_ expand>	The minimum limit of the wildcard patterns matching set from the SIETS storage vocabulary in absolute numbers. This parameter overcomes the cover_factor parameter. For example, if only 2 words fall in the cover_factor, but the min_exapand is 4, then 4 words are being used in the search.	4
	<max_ expand>	The maximum limit of the wildcard patterns matching set from the SIETS storage vocabulary in absolute numbers. This parameter overcomes the cover_factor parameter. For example, if 20 words fall in the cover_factor, but the max_exapand is 16, then only 16 words are being used in the search.	16

5.2.3.3. Stemming

The following table describes parameters for configuring stemming:

Third level element	Forth level element	Description	Default
<stemming>		This element contains parameters for configuring stemming.	
	<allow>	Information whether the language declinations search is enabled.	yes
	<cover_ factor>	When language declinations are used to define a class of words to be searched, only a limited number of statistically frequent words are searched for to ensure a higher performance. This element defines the limit in percent from the sum of all words created from the language declinations appearance in the SIETS storage. **Example:** Search query: car All words: car, cars, and car's. Number of times each word appears in the SIETS storage: 7 | 2 | 1 car car car car car car car | cars cars | car's Cover factor 80% means that only shadowed words are searched and returned. 10 20 30 40 50 60 70 80 | 90 100 car car car car car car car cars | cars car's Note that the word "cars" is searched, as it is important that at least one of the all appearances of "cars" fall in 80%.	95

Third level element	Forth level element	Description	Default
	<min_ expand>	The minimum limit of the language declinations matching set from the SIETS storage vocabulary in absolute numbers. This parameter overcomes the cover_factor parameter. For example, if only 2 words fall in the cover_factor, but the min_exapand is 4, then 4 words are being used in the search.	4
	<max_ expand>	The maximum limit of the language declinations matching set from the SIETS storage vocabulary in absolute numbers. This parameter overcomes the cover_factor parameter. For example, if 20 words fall in the cover_factor, but the max_exapand is 16, then only 16 words are being used in the search.	16

5.2.3.4. Alternatives Support

If the alternatives search is performed, the system returns a set of alternative words from the SIETS storage vocabulary, which are similar in spelling or has a different language declination, for example, if you enter "bote", then "bite" are "byte" are offered for searching. Note that only words from the SIETS storage are returned.

This feature can be used for fuzzy searches and for spelling error corrections.

The following table describes parameters for configuring alternatives support:

Third level element	Forth level element	Description	Default
<alternatives>		This element contains parameters for configuring alternatives support limits. When searching alternative words, the alternatives command considers the statistical information about the occurrence of the alternative word in the vocabulary, and the similarity of the alternative word to the search term. Although, the parameters for calculating alternatives similarity and appearance are defined when performing the alternatives command, the limit values for these parameters can be configured in the SIETS storage configuration file.	
	<cr>	Minimum ratio to include the alternative in the search query between the occurrence of the alternative and the occurrence of the search term. If you increase this parameter, there are less number of results returned to the result set, however performance is improved.	2.0

Third level element	Forth level element	Description	Default
	<idif>	Maximum number that indicates how much does the alternative differs from the search term, the greater the idif value, the greater the difference. If you increase this parameter, there are greater number of results returned to the result set, however performance is reduced.	3.0
	<h>	Minimum number that gives an overall estimation of the quality of the alternative, the greater the cr value and the smaller the idif value, the grater the h value. If you increase this parameter, there are less number of results returned to the result set, however performance is improved.	2.5

5.2.3.5. Ignored Words in Search Queries

The following table describes parameters for configuring ignored words options:

Third level element	Forth level element	Description	Default
<ignore>		This element contains parameters detecting ignored words.	
	<word_freq>	Ratio between all words in the SIETS storage and the word to be ignored. If this ratio for a word is less than this number, the word is added to the ignored word list.	500
	<word_len>	Maximum length of the word to be ignored.	5

Note: It is possible to include ignored words in the search by using the "+" sign in front of the ignored word. Full text index contains all words, including ignored words. The ignored words feature is used only for filtering out common words such as "and", "but", "is".

5.2.3.6. Example

The following is an example of the whole directory element:

```
<dictionary>
    <specsymbols>_</specsymbols
    <wildcards>
            <allow>yes</allow>
            <cover_factor>0.95</cover_factor>
            <min_expand>4</min_expand>
            <max_expand>16</max_expand>
    </wildcards>
    <national>
            <cover_factor>0.95</cover_factor>
            <min_expand>4</min_expand>
            <max_expand>16</max_expand>
    </national>
    <alternatives>
            <cr>2.0</cr>
            <idif>3.0</idif>
```

```
            <h>2.5</h>
    </alternatives>
    <ignore>
            <word_freq>500</word_freq>
            <word_len>5</word_len>
    </ignore>
</dictionary>
```

5.2.4. Repository

The following table describes SIETS storage repository configuration parameters:

Second level element	Third level element	Forth level element	Fifth level element	Description	Default
<repository>				This element contains the repository configuration parameters.	
	<highlight>			This element contains parameters for highlighting the matching search terms in the search result.	
		<open_ mark>		Highlight open mark.	
		<close_ mark>		Highlight close mark.	
	<snippet>			This element contains parameters for result list composition of search and lookup commands	
		<chunk>		Maximum number of bytes read from single document to compose result set.	16384

Example:

```
<repository>
    <highlight>
        <open_mark>&amp;lt;b&amp;gt;</open_mark>
        <close_mark>&amp;lt;/b&amp;gt;</close_mark>
    </highlight>
</repository>
```

5.2.5. Index

The following table describes SIETS storage indexing configuration parameters:

Second level element	Third level element	Forth level element	Description	Default
<index>			This element contains the indexing configuration parameters.	
	<cache>			

Second level element	Third level element	Forth level element	Description	Default
		<size>	Indexing cache size in mega bytes, from 50 to 150 MB. If you enter a number outside this interval, then: • If less than 50, the performance is very low. • If more than 150, the performance is not affected. Note that the indexing demon uses more RAM than this number, because there are also other operations. If you are importing a large data amounts in size of several GB, then the whole is being used.	80
		<usage_ idle>	Minimum indexing amount of the cache in percent. Only if this minimum is exceeded in the cache, the indexing is started. If the data amount in the cache is less than the minimum, the background indexing is not performed. Leave this parameter unchanged, unless advised by the SIETS technical support team.	10
		<usage_ critical>	Maximum indexing amount of the cache in percent. If the maximum is exceeded, all CPU and I/O resources will be used for indexing. If the data amount in the cache is less than the maximum, CPU and I/O resources for indexing are used proportionally the data amount in the cache. Leave this parameter unchanged, unless advised by the SIETS technical support team	90
	<background _indexing>		Information whether the background indexing is performed. If not, indexing is performed only when the index command is sent. Leave this parameter unchanged, unless advised by the SIETS technical support team	yes
	<optimize _to>		Number of search results to be optimized according to the relevance. Search results after this number are sorted by the rating. It is suggested to have this number the same as the maximum number of documents returned to the result set. The greater the number, the more relevant search results. The lesser the number, the higher performance.	1000

Second level element	Third level element	Forth level element	Description	Default
	<weight _threshold>		Weight threshold for the relevance, which is considered as a very relevant. For example, if 100 is the maximum relevance weight interval, then 90 is very close to the maximum, but also is likely that documents with such relevance exists in reality. Therefore, it is considered as very relevant.	90

Example:

```
<index>
    <cache>
        <size>80</size>
        <usage_idle>10</usage_idle>
        <usage_critical>90</usage_critical>
    </cache>
    <background_indexing>yes</background_indexing>
    <optimize_to>1000<optimize_to>
    <weight_threshold>90<weight_threshold>
</index>
```

6. RUNNING SIETS COMMANDS

SIETS Enterprise Manager contains additional functionality beyond administering and configuring SIETS environment and SIETS storages. This functionality usually is performed using SIETS API, however, it is also implemented in SIETS Enterprise Manager for testing purposes.

After you have set up the SIETS environment and configured the SIETS storage, you can run a SIETS command to test how the system works.

To run a SIETS command from SIETS Enterprise Manager, proceed as follows:

1. After you have logged in SIETS Enterprise Manager, in **Main Menu,** select **SIETS Storages.**
2. In the **SIETS Storages** window, in the **Name** column, select the SIETS storage you want to run the SIETS command for.
3. Select **SIETS Command** below the instance.
4. To configure an instance, select **Configuration** below the instance.

   The **SIETS Command** window appears.

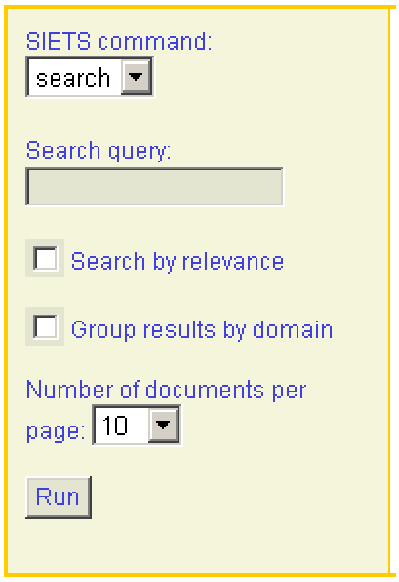


Figure 22: Running SIETS command

5. In the **SIETS command** drop-down list box, select the command to run.
6. Enter the SIETS command parameters according to the selected command in fields described in the following table:

SIETS command	Parameter	Description
search		Performs FTS in the SIETS storage.
	Search query	Search query.
	Search by relevance	Information whether the results are ordered by relevance.
	Group results by domain	Information whether the results are ordered by domain.

SIETS command	Parameter	Description
	Number of documents per page	Number of documents displayed on one result page.
retrieve		Returns a document from the SIETS storage. If a document with such ID is not in the SIETS storage, the function returns an error.
	Document id	Document ID to be retrieved.
lookup		Searches for a document in the SIETS storage and returns the information whether the document with such ID exists is in the SIETS storage or it does not.
	Document id	Document ID to be looked up.
status		Returns status information of the SIETS server instance, which includes the following: • number of documents in the SIETS storage • number of words in the vocabulary • total number of words in the SIETS storage • number of executed commands since the last startup of the instance • number of errors that have occurred since the last startup of the instance
index		Tells the SIETS server to start the process of indexing.
clear		Deletes all documents from the SIETS storage. This function should be used only when a complete re-indexing of the SIETS storage is necessary.
delete		Deletes a document from the SIETS storage.
	Document id	Document id to be deleted.

For more information on SIETS commands, see the *SIETS Developer's Guide.*

7. To run the SIETS command, select **Run.**

APPENDIX A: FILE STRUCTURE

This appendix describes the default file structure of the SIETS system.

After the SIETS server installation, the siets directory is created under usr/local:

usr/local/siets.

In the siets directory, there is the data directory that contains all information including data, configuration, and log files for each of the SIETS storages added to the SIETS server in a separate directory named after the SIETS storage.

For example, all information of the test SIETS storage is located in usr/local/siets/data/test.

The SIETS storage directory contains the following files:

Title	Description
config.xml	SIETS storage configuration file.
DIC	Prefix for data files names that contain vocabulary.
DOC	Prefix for data files names that contain repository data.
MTX	Prefix for data files names that contain inverted index.
log_YY_MM_DD	Access log file for the date appearing in the file name. Note that all log files are automatically deleted periodically as they can increase in size. For information on configuring the period in which log files are deleted, see SIETS Storage Configuration File Parameters, General.
error_YY_MM_DD	Error log file for the date appearing in the file name. Note that all log files are automatically deleted periodically as they can increase in size. For information on configuring the period in which log files are deleted, see SIETS Storage Configuration File Parameters, General.
parsing_xml_YY_MM_DD	Request X ML parsing log file. Note that XML parsing errors cannot be added to the error log, because, before parsing the XML request, it is not possible to retrieve information for construing the log, for example the request ID.
dump_YY_MM_DD	Dump of commands that are sent to the SIETS server. This log is created only if the SIETS storage is configured respectively. For information on turning on or off the dump log see SIETS Storage Configuration File Parameters, General. Note that all log files are automatically deleted periodically as they can increase in size. For information on configuring the period in which log files are deleted, see SIETS Storage Configuration File Parameters, General.

APPENDIX B: SIETS ADMINISTERING SHELL COMMANDS

This appendix contains several SIETS administration shell commands for administrators that prefer working from a shell than using the SIETS Enterprise Manager web interface.

You can configure the SIETS storage in the textual mode by editing the config.xml file of the SIETS storage in any text redactor, for example, vi, pico, emacs.

For more information on the SIETS file system, see Appendix A: File Structure.

The following table contains commands for reviewing logs using the shell:

Command title	Description	Example
tail	Opens the end of the log.	tail log_03_11_05
less	Opens the whole log.	less log_03_11_05
grep	Filters a part of the log records. It can be used to filter records of specific commands, IP addresses, and so on.	cat log_03_11_05 | grep siets_command:search | less cat log_03_11_05 | grep 195.244.157.87 | less The cat command prints the whole log, then redirects the output to the grep commend, which filters only those records that contains the string 'siets_command:search', or 195.244.157.87 in the second case.

For more information on the shell commands, use the man command, for example, man less.

GLOSSARY

A	
API	Application programming interface.
D	
demon	Program or process, part of a larger program or process, that is dormant until a certain condition occurs and then is initiated to do its processing
document repository	Place where all SIETS documents are stored in the format, in which they were stored in the SIETS system, for returning the documents on a search request. See also: SIETS document.
F	
FTS	Full text search. See also: FTS query.
FTS query	Full text search request to the SIETS server. See also: FTS.
I	
inverted index	List of words, where each word has a list of pointers to SIETS documents in which the word occurs. See also: SIETS document.
R	
RAM	Random access memory.
rate	Mechanism, which ensures that results are ordered in a result set according to assigned unchanging rate.
relevance	Measure of the accuracy of the search results. See also: rate.
S	
scheme	Document structure definition mechanism, which defines the location and behavior of each document part.
SIETS API	Standardized set of functions for accessing the SIETS server. See also: SIETS server.
SIETS Enterprise Manager	Administering tool of the SIETS system.
SIETS console	Simple text application for accessing the SIETS server directly using the same functions as in SIETS API. See also: SIETS API.
SIETS document	Unit in SIETS storage against which searching is performed. It can be unstructured or XML structured. See also: SIETS storage.
SIETS server	Stand-alone server for storing and retrieving information such as plain texts or XML structured documents. It can be run in one or more instances per computer.
SIETS storage	Data collection for storing SIETS documents in a format that ensures a search is performed very fast. SIETS storage is contained by one SIETS server instance, and consists of vocabulary, document repository, and inverted index. See also: SIETS document, vocabulary, document repository, inverted index.

stemming	Feature that allows searching for words and their declinations.
U	
UTF	UCS (universal character set) transformation format.
V	
vocabulary	Vocabulary is a list of all unique words in the SIETS storage. Unique words are found in documents and added to the vocabulary while storing these documents to the SIETS storage. See also: SIETS storage.
W	
wildcard search	Feature that allows searching for unknown characters or phrases.
X	
XML	Extensible markup language.
XML reply	XML message that is returned when submitting a XML request. See also: XML request.
XML request	XML message that is sent to the SIETS server to perform a SIETS API command. See also: SIETS API.

INDEX