



HULBEE ENTERPRISE SEARCH

Installation and setup manual

Modified: 27-Jan-2017

Version: 2.2.45

Table of Contents

USED TERMS	3
1 ARCHITECTURE	4
1.1 HULBEE ENTERPRISE SEARCH.....	4
1.2 HESBOX.....	5
2 INITIAL INSTALLATION	5
2.1 FULL INSTALLATION	5
2.1.1 Indexstore	5
2.1.2 Processing server	6
2.1.2.1 Hardware requirements for HES Processing Server	6
2.1.2.2 Software requirements for HES Processing Server	6
2.1.2.3 IFilters setup.....	7
2.1.2.4 Authentication settings	7
2.1.2.5 HES unpacking.....	7
2.1.2.6 IndexUtil and custom index structure	8
2.1.2.7 Initial setup.....	9
2.1.2.8 License setup.....	13
2.1.2.9 Desktop Manager setup and settings.....	13
2.2 HESBOX.....	14
2.2.1 Domain setting up	14
2.2.2 File storage	15
3 FIRST RUN	15
3.1 INDEX CLEANING AND MODIFYING	16
4 ADMIN PAGE SETUP	17
4.1 RESTORE OPTIONS.....	17
4.2 THE FIRST SETUP STEPS.....	17
4.3 CONNECTOR REGISTRATION	17
4.4 ADDING A STORAGE OF DOCUMENTS.....	18
4.5 ADMINISTRATION AREA	18
4.5.1 Dashboard	19
4.5.2 Connectors.....	20
4.5.3 Users	20
4.5.4 Logs.....	21
4.5.5 Unprocessed	21
4.6 SETTINGS.....	22
4.6.1 Authorization	22
4.6.2 Processing	24
4.6.3 Cleaning.....	27
4.6.4 Logging	28
4.6.5 Import / Export	29
4.6.6 Preview	29
4.7 DATA CLOUD.....	30
4.7.1 Custom queries	31
4.7.2 Stopwords.....	31
5 CONNECTORS	32

5.1	COMMON TUNES	32
5.1.1	<i>Settings of the connectors via the configuration files.....</i>	32
5.1.2	<i>Settings of the connectors in the admin area.....</i>	34
5.1.3	<i>Manual connector setup.....</i>	37
5.1.4	<i>Connector unregistration.....</i>	37
5.2	FILESYSTEM CONNECTOR	38
5.2.1	<i>File connector settings.....</i>	38
5.3	MS-EXCHANGE CONNECTOR.....	39
5.3.1	<i>MS-Exchange Connector settings.....</i>	39
5.4	WEB CONNECTOR.....	39
5.4.1	<i>Settings of the Web Connector.....</i>	40
5.4.2	<i>Fine tuning of the Web Connector.....</i>	42
5.4.3	<i>Disable the indexing of a part of a web page.....</i>	43
5.4.4	<i>Canonical links.....</i>	43
5.4.5	<i>X-HES-Users headers.....</i>	43
6	USEFUL LINKS	44
7	KNOWN ISSUES	45

Used terms

Index – related to HES a data structure containing documents and additional information. This information allows you to find quickly the document you are looking for using the search by terms. It is also used to store HES settings.

Storage – represents a physical storage with documents, like a file system, mail service, “network disk”, “ftp server”, “some site”, etc.

Feed – list of metadata for documents from storage that should be processed. One feed represents one storage.

Connector – a plugin that allows HES to search the documents stored in different custom repositories such as enterprise file system, web resources, mail service for further processing. A connector is a REST-service application providing methods (endpoint) that the Connector Manager calls to get documents from certain data sources.

Crawler – part of the Web Connector that traverses the pages in the local and global network. Crawler has its own settings and period of crawling, which are installed when adding storage. Unregistering of the connector does not stop the crawler.

1 Architecture

1.1 Hulbee Enterprise Search

Hulbee Enterprise Search (HES)¹ is designed to work with MS Active Directory users and to conduct a search of the different types of files in different data sources. Since Version HES 2.1, in addition to the enterprise file system, the other types of resources are available “from the box”, e.g. web resources, MS-Exchange resources, etc. They are connected to the system via the connectors. In addition, there is the API, which allows developing custom connectors. All sources that are available for searching, as well as individual storages within the sources, are added to the system by administrator of the company.

Users’ file permissions are also taken into consideration. Users should at least have a document read permission to see files in the search results.

Supported file formats²:

File type	Extensions	Text extraction	Meta tags extraction	Attached files extraction
Text	txt, rtf, doc/dot, odt, wri, sxw	✓		
	docx/docm/dotx	✓	✓	
Publication	Pdf	✓	✓	
	Xps	✓		
Hypertext	html, htm, xml	✓		
	mht, shtml	✓	✓	
Tables	xsl, xslt, xls, ods, csv	✓		
	xlsx	✓	✓	
Presentations	pptx	✓	✓	
	ppt, pps, odp	✓		
Graphics	bmp, jpg/jpeg, png, jfif, tif, tiff, jpe		✓	
E-mail	msg, eml	✓	✓	✓
Archives	zip, rar, 7zip			✓
Media	avi, mp3, mp4, wav, m4a, wma, wmv,ogg, flac, mkv, ape, mpc		✓	
Source Code and Scripting	cs, vb, js, csproj, h, c, cpp, vbs,vcproj, vbproj, pl, sql, bat, cmd	✓		
	css	✓	✓	

¹ Some of the modules have Swisscows or SES in the filename or texts. It is the old name of the project and is synonymous for Hulbee Enterprise Search or HES.

² The possibility of the search system is not only to index file metadata (file name, path, size, creation date, modification date), but also to work with its content – to extract the text and/or meta tags, and/or unpack files, containing other files.

Any modern browser (Mozilla Firefox, Chrome or Internet Explorer latest versions) with the opened link to the search engine in intranet can be used as a user interface. The search and user settings modification are performed here. Using admin panel, administrators can also adjust various settings.

Documents that are available as files in a file system (network resources), HES is trying to open exactly as files, using the program which is associated with their extension (for example, Microsoft Word or Acrobat Reader), but does not download them from the browser. To make this possible, on the user's computer, you must be running the module Desktop Manager, which is described in detail in the User Manual.

To open search-found files, a user needs to run a computer under his/her own real account in one local network with file storage. A user needs to possess the necessary rights for software installation or to turn to the administrator for help.

HES software complex consists of two main parts:

- Indexstore (Linux server with installed and configured Elasticsearch).
- Processing server (Windows server with all other components).

In the case of small filestores, these servers might be combined into the single Windows Server Machine. Large filestores may require mounting the Elasticsearch cluster containing a number of servers.

1.2 HESbox

The present configuration is the appliance server, containing all necessary components of Hulbee Enterprise Search. Taking into consideration the fact that it contains the full-featured Windows Server 2012 R2 Standard, it can be used not only for search but also for deployment of services based on MS Active Directory³ and storage of data in medium-sized enterprises⁴, which still use peer networks.

HESbox contains versions notable both for Hardware capabilities and for limitation of a number of users.

2 Initial Installation

2.1 Full installation

2.1.1 Indexstore

Elasticsearch v.2.4.* is used as an indexstore. Install it, using directions from the manufacturer website (see chapter 6). Elasticsearch could be installed both on the computer, containing Processing server components, and on individual computer. If you install Elasticsearch on individual computer, you may use an operating system different from Windows, but supporting JRE (GNU/Linux, Solaris, etc).

In the case of extremely high loads, it is recommended to use a cluster containing a number of Elasticsearch servers.

³ You may find links to introductory articles about Active Directory in chapter 6 (Useful links).

⁴ Remember that you need to adjust and regularly back up your data. You may find the general information about data back-up in chapter 6 (Useful links).

After installation in accordance with Elastic website recommendations, it is necessary to make post-installation settings in Elasticsearch configuration (elasticsearch.yml file). It is desirable to name the server cluster Elasticsearch other than the default (cluster.name key). This is necessary to minimize the risk of index damage or compromise when deploying other instances of Elasticsearch on the local network.

Notice! Index may contain confidential data. To prevent leaks of such data please disallow all TCP/IP connections for all of the components except Application Server.

The performance of Elasticsearch is significantly affected by the performance of the disk subsystem and the amount of memory available for buffers Elasticsearch and file cache. The optimal amount of memory for buffers Elasticsearch depends on many parameters, including the nature of the data, but as initial values for the HES, we can recommend this algorithm:

For buffer Elasticsearch it is necessary to allocate 25% of RAM, if you install Indexstore on a separate machine and 20% of RAM if you install on a single machine with Processing server. But this number should be at least 3GB and not more than 31GB.

In Linux, the amount of memory allocated to ES, is set as the environment variable ES_HEAP_SIZE. When you install Elasticsearch on Windows as a service, the amount of memory can be set via the “Elasticsearch properties” GUI. It will be available after running command “service.bat manager” through the command line.

2.1.2 Processing server

2.1.2.1 Hardware requirements for HES Processing Server

Component	Minimum	Recommended
Processor Cores	4	>=8
Memory	16 GB	64 GB
Hard disks and available storage space	256 GB	512 GB
Network adapter speed (to filestorage and indexstorage)	1 Gb/s	>=10 Gb/s

2.1.2.2 Software requirements for HES Processing Server

Install Windows Server 2012 R2 Standard with the latest updates and the following components:

NetFx4ServerFeatures	IIS-RequestFiltering	IIS-ISAPIExtensions	IIS-WebServerManagementTools
NetFx4	IIS-StaticContent	IIS-ISAPIFilter	
NetFx4Extended-ASPNET45	IIS-DefaultDocument	IIS-ASPNET45	IIS-ManagementConsole
IIS-WebServerRole	IIS-DirectoryBrowsing	IIS-HealthAndDiagnostics	WCF-Services45
IIS-WebServer	IIS-HttpErrors	IIS-HttpLogging	WCF-TCP-PortSharing45
IIS-CommonHttpFeatures	IIS-ApplicationDevelopment	IIS-Performance	IIS-WebSockets
IIS-Security	IIS-NetFxExtensibility45	IIS-HttpCompressionStatic	
		IIS-WindowsAuthentication	

Before installation, you can execute the following command:

```
Dism /Online /Enable-Feature /FeatureName:NetFx4ServerFeatures  
/FeatureName:NetFx4 /FeatureName:NetFx4Extended-ASPNET45  
/FeatureName:IIS-WebServerRole /FeatureName:IIS-WebServer  
/FeatureName:IIS-CommonHttpFeatures /FeatureName:IIS-Security  
/featurename:IIS-WebSockets /FeatureName:IIS-RequestFiltering  
/FeatureName:IIS-StaticContent /FeatureName:IIS-DefaultDocument  
/FeatureName:IIS-DirectoryBrowsing /FeatureName:IIS-HttpErrors  
/FeatureName:IIS-ApplicationDevelopment /FeatureName:IIS-  
NetFxExtensibility45 /FeatureName:IIS-ISAPIExtensions /FeatureName:IIS-  
ISAPIFilter /FeatureName:IIS-ASPNET45 /FeatureName:IIS-  
HealthAndDiagnostics /FeatureName:IIS-HttpLogging /FeatureName:IIS-  
Performance /FeatureName:IIS-HttpCompressionStatic /FeatureName:IIS-  
WebServerManagementTools /FeatureName:IIS-ManagementConsole  
/FeatureName:WCF-Services45 /FeatureName:WCF-TCP-PortSharing45  
/FeatureName:IIS-WindowsAuthentication /All
```

You may also do it, using system applet “Turn Windows features on or off”.

2.1.2.3 IFilters setup

Install the following IFilters to get more exact text extraction from MS Office and PDF documents:

- MS Office: <http://www.microsoft.com/en-US/download/details.aspx?id=17062> with service pack <http://support.microsoft.com/kb/2687447>. Install 64-bit versions.
- PDF IFilter 64 11.0.01: <http://www.adobe.com/support/downloads/detail.jsp?ftpID=5542>.

2.1.2.4 Authentication settings

HES Installer installs WEB-section (user interface and admin panel) as an Application to the IIS Default Web Site. The administrator can change the settings of the application using the standard tools of IIS and Windows administration.

The Auto-login to the HES system, which is used to implement fast authentication without entering a username and password of the current Windows user, requires the inclusion of the appropriate IIS configuration:

1. Select a site with HES (usually it is the Default Web Site) in the IIS Manager window.
2. Select “Authentication” in the features panel. Here it is necessary to include the following items: “Anonymous Authentication” and “Windows Authentication”.
3. Choose the appropriate “Negotiate” provider for “Windows Authentication” in the section “Action”> “Enable”> “Providers ...”.

2.1.2.5 HES unpacking

HES applications pack has a name like **HES.2.2.XX.XXXXX.zip** (XX being the numbers of your specific version). It contains the following components:

- Connectors
 - Hes.Connectors.Exchange
 - Hes.Connectors.FileSystem
 - Hes.Connectors.Web

- Helpers
 - ConfigTransformationHelper
- Services
 - Hes.Services.ConnectorManager
 - Hes.Services.IndexCleaner
- Utilities
 - HESCoreMock
 - IndexUtil
- Web
 - Ses.Web
- Ses.Setup.*

Unpack it to the some folder, for example C:\HES (i.e. C:\HES\Web\, C:\HES\Services\, C:\HES\Utilities\, etc).

Included with HES is also the Desktop Manager utility to open files from the enterprise file system. For more information on the installation, see section 2.1.2.9.

2.1.2.6 *IndexUtil and custom index structure*

During subsequent installation, the installer will create an index with default settings that should be used in most cases. However, there are cases when you need to create not the usual configuration. Then, we can previously create it using a utility IndexUtil (distribution in the HES). The installer, on the stage of the task of the index to use, will determine that the index has already been created and will be offered a choice to re-create it or use what is already created.

HES uses 3 Indices:

Index name	Content	Default parameters
{HES}	Text content for search.	Shards = 4 Replicas = 1
{HES}_settings	Settings	Shards = shards in HES index Replicas = 1
{HES}_diagnostics	Logs of the HES core	Shards = shards in HES index Replicas = 1

In this case, {HES} is the index name specified by the utility IndexUtil as a parameter or in the installation process. This name is entered in the “Index Name” field in Step 2 of the installation (see. 2.1.2.7). Name of the subsidiary indices is constructed by means of suffixes added to this name. To get information about the settings utility, start it with the following key:

```
IndexUtil.exe /?
```

Sets the number of shards & replicas other than the default, it makes sense, when a cluster of Elasticsearch with the number of nodes is not equal to 2 or 4. Recommendations regarding the selection of these parameters can be found on the developer’s website Elasticsearch (see 6 – Useful links).

Notice! If Elasticsearch has multiple nodes, avoid changing any settings in admin panel during the restart of a server.

2.1.2.7 Initial setup

Run the ses.setup.exe utility in the root folder of unpacked distributive.

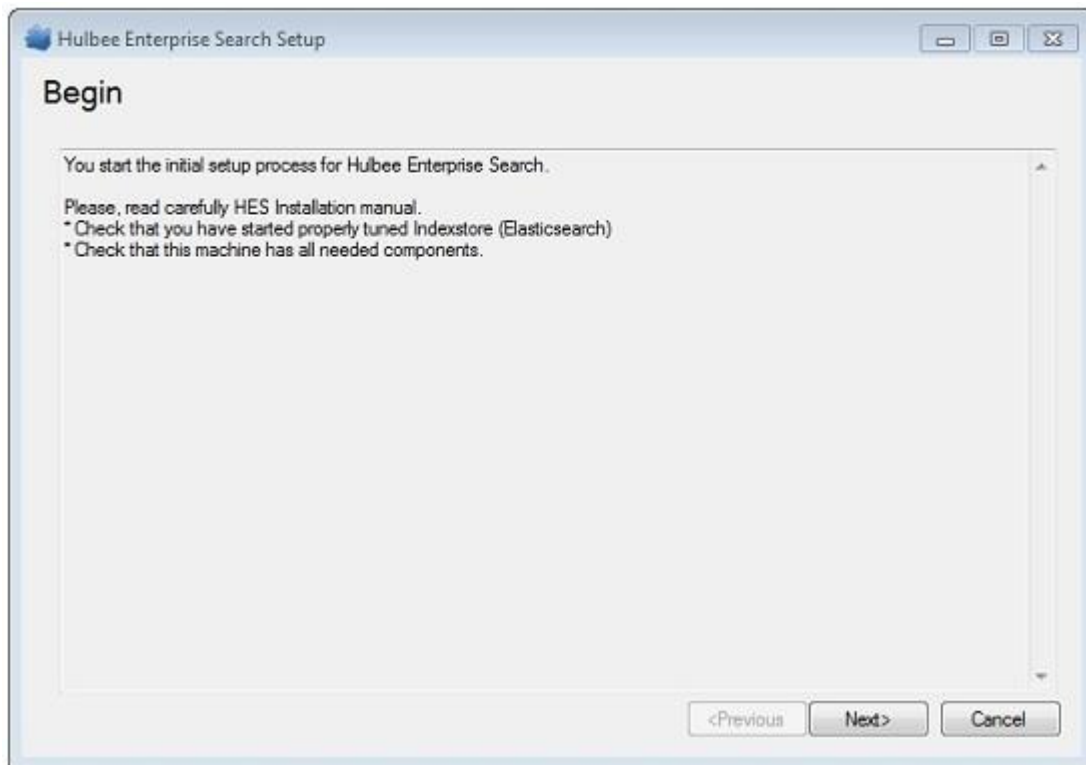


Fig. 1. HES Setup – start of Setup.

Click the "Next" button to take step 1.

1. Step 1.

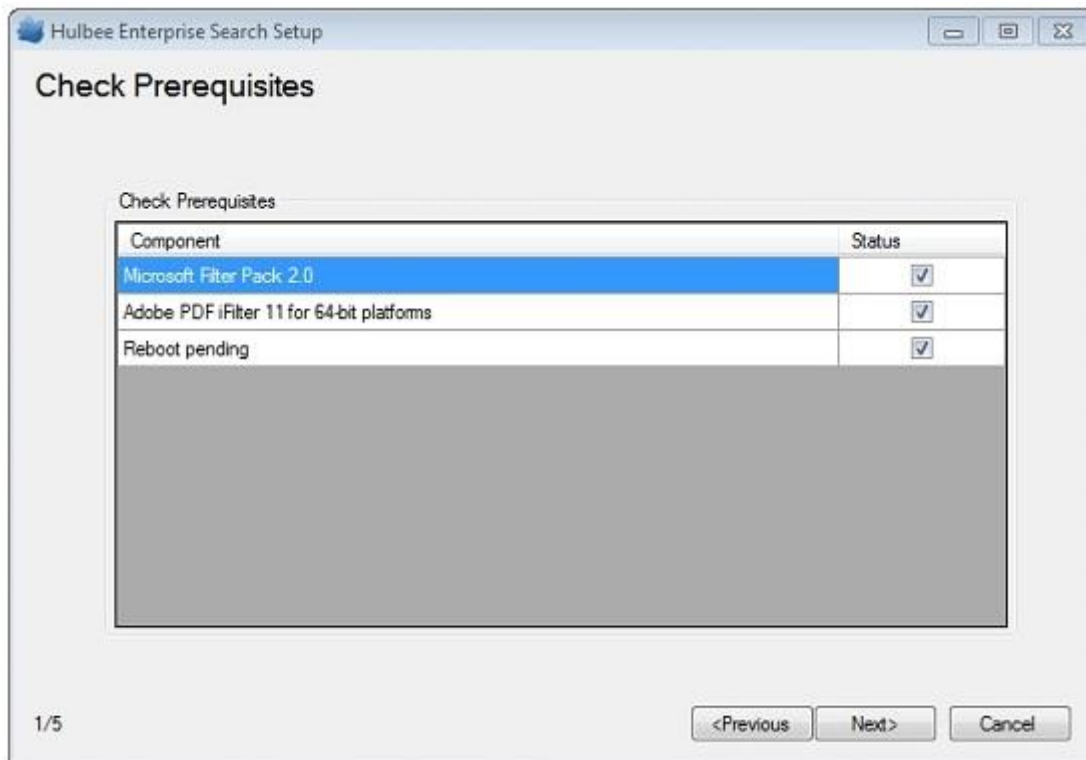


Fig. 2. Step 1 of HES Setup.

At the stage of optional prerequisites check, missing components should be specified. The setup can be continued, but the lack of IFilter could have an impact on the quality of extraction of the text from some formats (MS Office and PDF).

2. Step 2.

Enter the valid URI to the Elasticsearch server and the index name (default value is "hes"). URI should have http protocol prefix, proper IP or domain name and port.

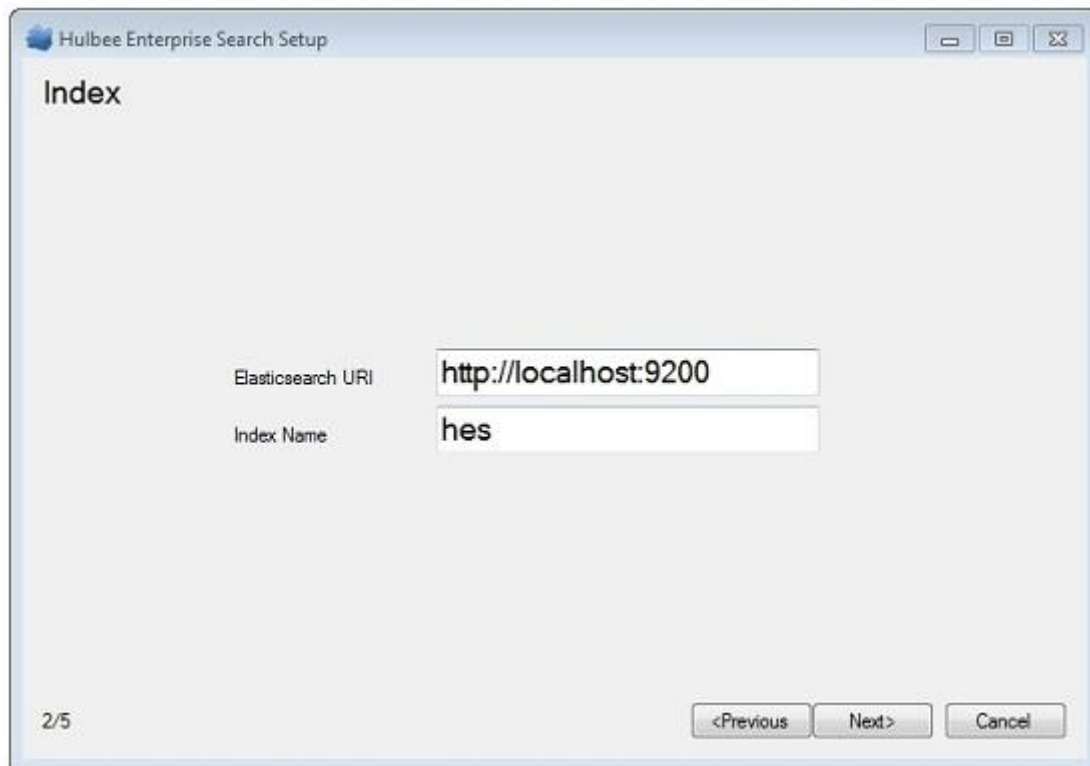


Fig. 3. Step 2 of HES Setup.

3. Step 3.

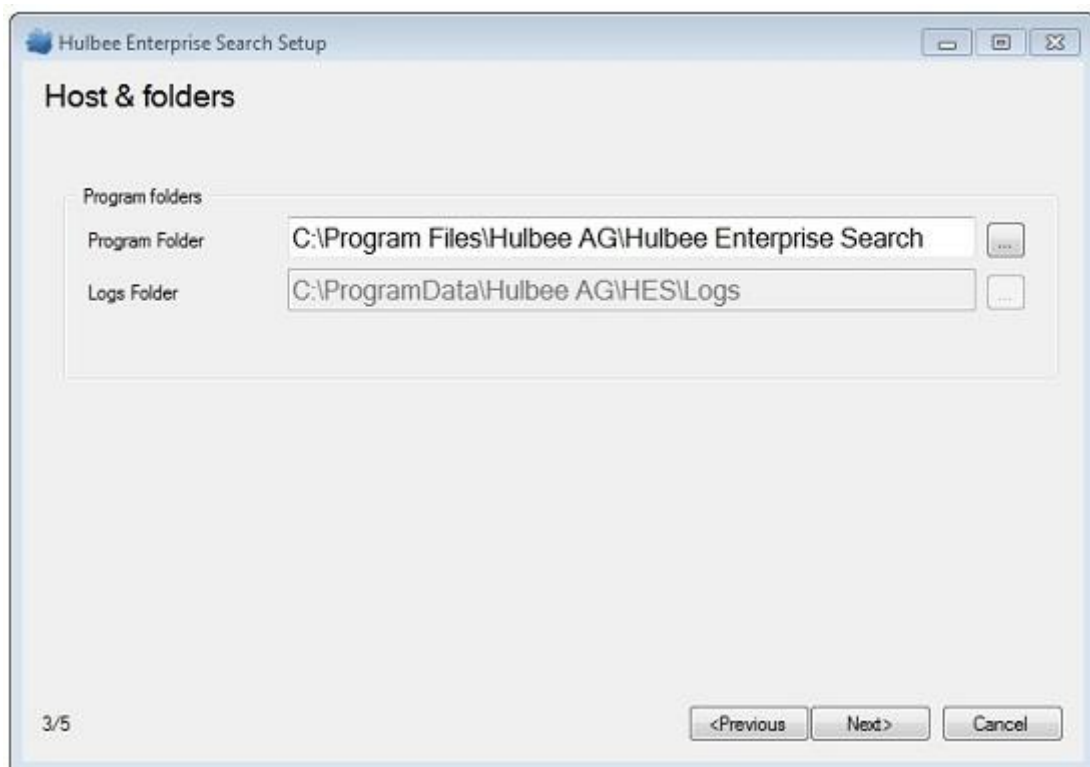


Fig. 4. Step 3 of HES Setup.

"Program Folder" and **"Logs Folder"** can be left with default values.

4. Step 4.

In this step, the HES administrator password is entered. This user is very special – he is not a user of the Active Directory or the current machine and he is needed in order to be able to authenticate on the HES software that is not set up yet. Only he has access to the admin panel of HES and can make administrative settings, such as connection to the Active Directory.

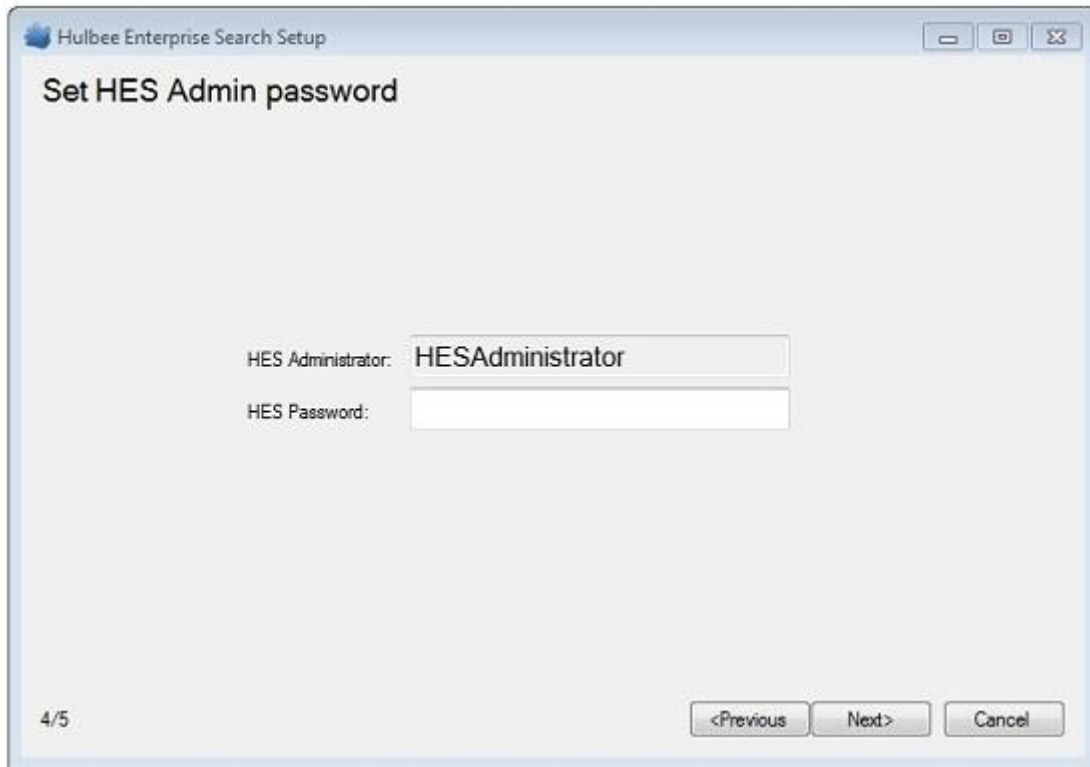


Fig. 5. Step 4 of HES Setup.

5. Step 5.

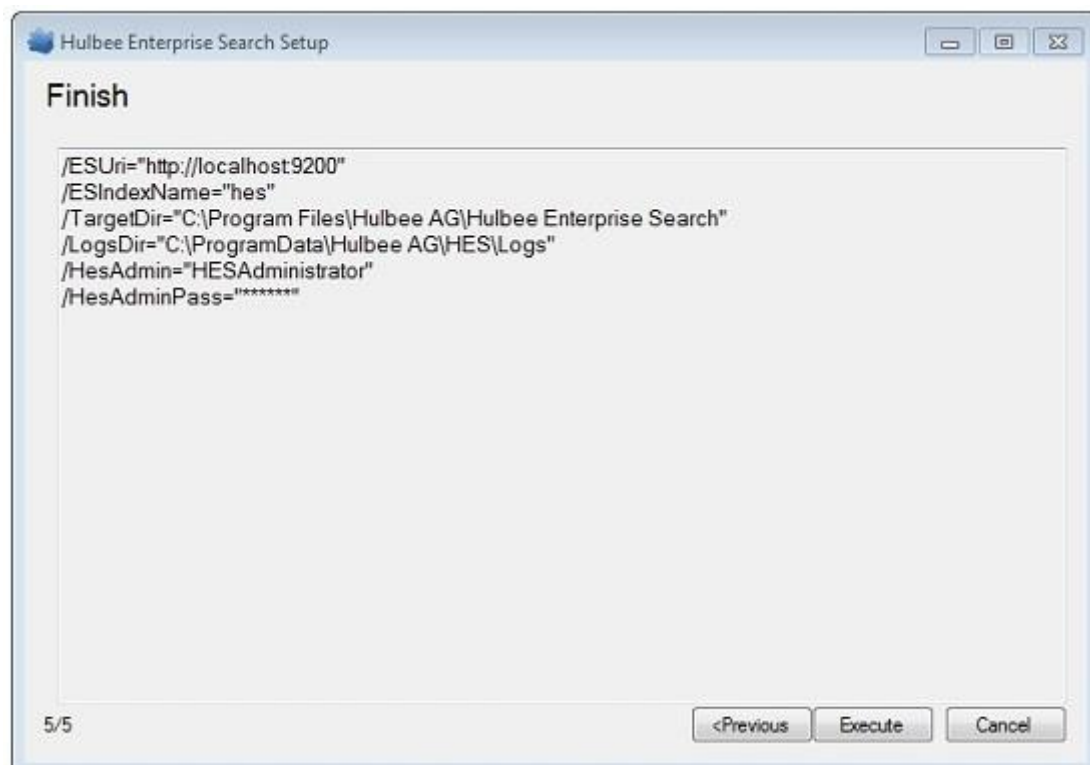


Fig. 6. Step 5 of HES Setup.

You can see the selected options and run setup with “Execute” button.

6. Final step.

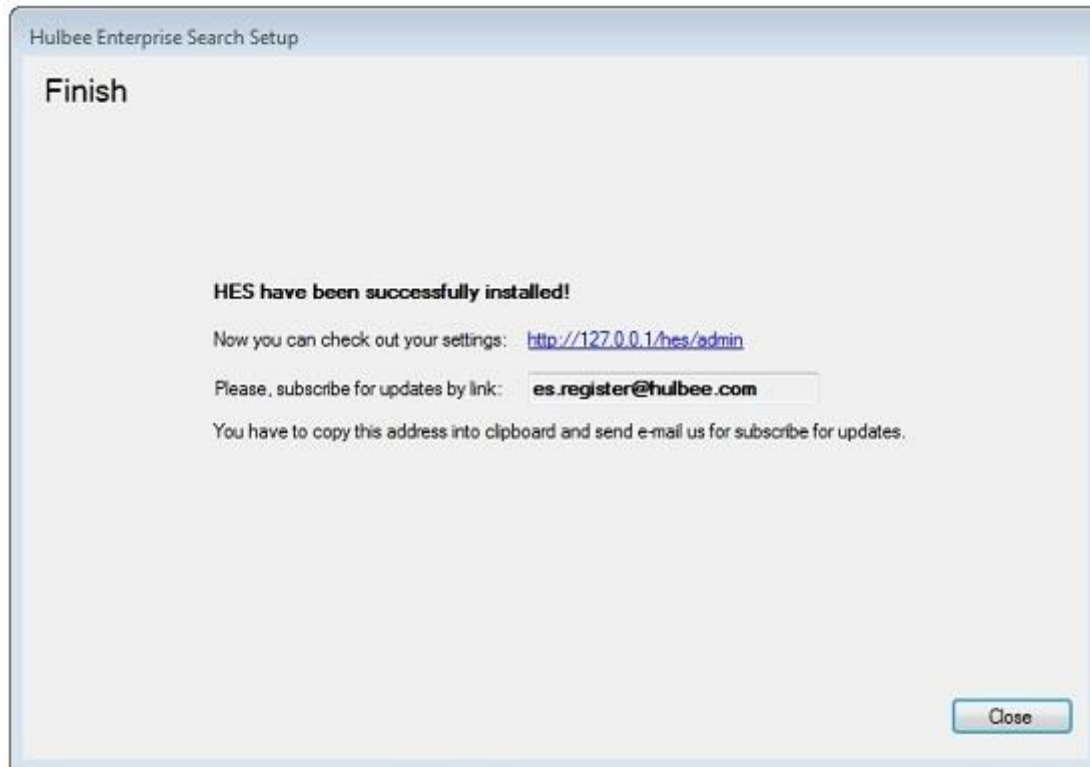


Fig. 7. HES Setup – setup finish.

After successful setup, it is necessary to visit the provided link (Fig. 7) with the same name and password that you used in step 4 of the installation, to perform basic configuration (section “Authorization” (see 4.6.1)) and register connectors (see 4.3).

2.1.2.8 License setup

In addition to the product, a personal license file is provided. If the file was not made available, you could request it from customer support support@hulbee.com. For the placement of the license file, the following folder is recommended: C:\Program Files\Hulbee AG\Hulbee Enterprise Search\Web\Ses.Web\bin (the same place where the binary modules of the HES site are installed). If the file does not exist, users from the Active Directory of the company will not be allowed for the use of the product.

2.1.2.9 Desktop Manager setup and settings

Utility Desktop Manager is only used to open files from the file storage of the company or the folder containing the files, if opening the file is not possible. It provides an opportunity to work with the original file and not with its copy, downloaded via a browser. This utility is described in the User Manual (Section 7.2).

After installing the HES you can find the file Hes.Desktop.Manager.Setup.msi in the installation folder, namely Web\Ses.Web\Downloads\.

Desktop Manager should be installed and started on the client computer. There are two installation scenarios: individually for each user and for all users of the client computer. The second installation option is executed only if you have administrator rights. Only one scenario can be executed on the same machine at the same time. After installation, the utility is added to the autorun.

The user can install Desktop Manager on their own via the HES website. The installation suggestion appears when the user tries to open a file from the enterprise file system if the current version of the utility is not already running. During installation, the user will be asked to enter the HES URL.

To simplify the installation for the domain users, install centrally on all computers of the domain the following registry key:

HKEY_LOCAL_MACHINE\SOFTWARE\Hulbee AG\HES

The string value “URL”, where the URL for installing HES is specified. If the Desktop Manager installation program finds such a sequence value, it skips HES URL configuration dialog and uses the value from the registry.

To change the settings, you can click the “Options” Desktop Manager context menu. By default, “Prefer HES URL from registry” is enabled. If necessary, turn off this setting and add HES URL manually. For the “for all users” scenario, the “Options” menu will be available, only when utility run with the administration rights.

Notice! Perform all the necessary settings to prevent the possibility of improper Desktop Manager settings on individual users’ computers.

If security policy of your domain prohibits users from installing on their machines arbitrary software, add the certificate from the distribution to the list of allowed to be installed.

2.2 HESbox

This configuration comes with all pre-installed components, necessary for work of HES and installed but not configured HES.

2.2.1 Domain setting up

First, plan the appliance server usage scenario. Depending on it, the following standard scenarios are available:

- Joining the HESbox Server to an existing domain.
- Creation of a new domain based on Microsoft Active Directory in Windows Server 2012 R2.
- Using the impersonation mode, without the introduction of a computer in the Active Directory domain.

Clicking the links in chapter 6 (Useful links), you will find a useful background information on the work with Active Directory.

In the process of distribution, administrator may sign in to the server, using pre-installed Administrator account with password “Admin123”.

Notice! Be sure you change the administrator password before the actual use of the server.

To change Windows administrator password use the standard procedure.

To change the HES administrator password take the following steps:

1. Open the configuration file C:\Program Files\Hulbee AG\Hulbee Enterprise Search\Web\Ses.Web\Web.config
2. Change the password at <add key="SuperUser.Password" value="Admin123" />
3. Save file.

2.2.2 File storage

The typical size of search index is up to 10% of binary document size (approximately, as it essentially depends on real database set). So, the appliance server may also be used as network file storage.

For example, if you suppose that within several years the file volume will not exceed 1 TB, and your configuration contains 2 TB of disk space, it is quite safe to organize not only appliance server based search, but also file storage. This scenario will also decrease network load during intensive indexing.

If you suppose that volumes will be higher, you need to store files on the server with sufficient storage capacity disk array or to set additional disks in appliance server.

Notice! The file storage should be available at the local network and should use the same Active Directory as other HES parts.

3 First run

The address, at which the installed HES software is available, depends on the administrative setup of the local network and to which site in IIS Application HES is attached.

If the opening is carried out from a local computer, it is usually:

<http://127.0.0.1/hes> or <http://localhost/hes>.

From other computers of the local network HES will be available either by the IP of the computer or by its name in the local network:

<http://hes-server/hes>.

After following the link, fill in an authorization form:

The image shows a web-based authorization form titled "Authorization" in a blue header. Below the header, there are two login options separated by "or": "Login using Windows account" and "Login using username and password". Under "Login using Windows account" is a large blue square icon with the white Windows logo. Under "Login using username and password" are two input fields: "User name" with a person icon and "Password" with a key icon. Below these fields is a green "Sign in" button with a gear icon to its right.

Fig. 8. Authorization form.

Notice! To administer the HES, you must enter an administrator name/password that were specified during installation in Step 4, or those for which they have been changed (see 2.2.1).

After the login the start page will appear:

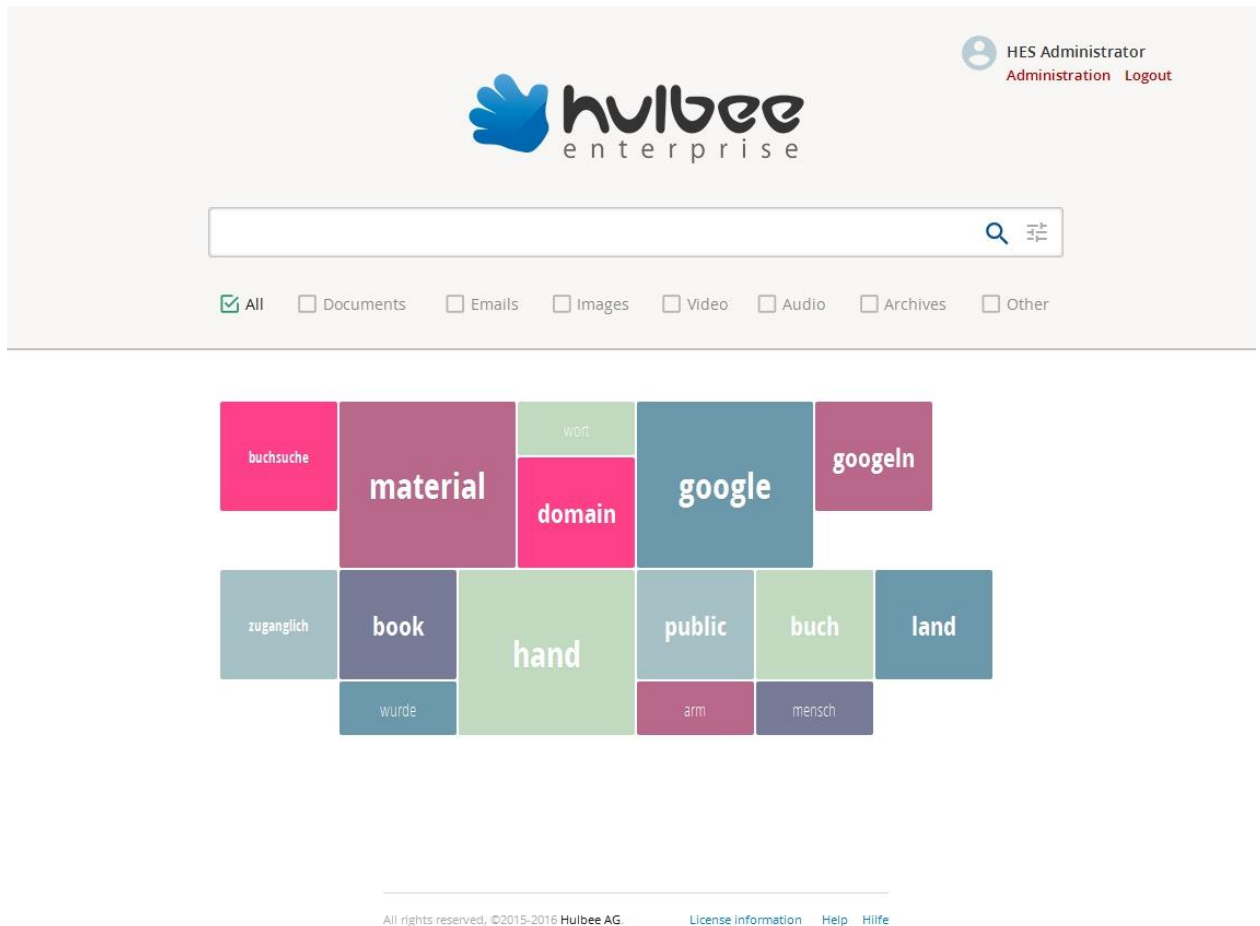


Fig. 9. HES start page.

As long as the index is empty, the DataCloud will not be displayed on the home page. This is a normal behavior of the system.

To access the Admin area, follow the link “Administration”.

3.1 Index cleaning and modifying

If you need to modify the structure of the index, or just clean it (with the command re-create or delete the index), you can use the utility IndexUtil. Using IndexUtil the index can also be re-created in the event that it was damaged.

Notice! Before working on the index, stop all the services and connectors of HES in the system Services applet.

Using IndexUtil for these operations is similar to described in Section 2.1.2.6. After recreating, the index (including settings) will be cleared and the system is shown in a state similar to that which existed after

the installation (except the settings stored in the configuration files). Accordingly, the adjustment should be introduced again as described in section 4.

After rebuilding the index, start the HES services again.

4 Admin page setup

The remaining setup steps can be done using admin panel. Please follow the next link:

<http://enterprise-search.company.com/hes/admin>

(instead of **enterprise-search.company.com** enter the domain, used in chapter 3) or follow the “Administration” link on the HES start page.

4.1 Restore options

If you have already installed and tuned instance of HES, you can restore previously backed up settings. Select “Import/Export” (see 4.6.5) area and:

1. Click “Browse...” button and select the back-up file.
2. Click “Import” button.

This operation allows restoring global and personal users settings. In other words, visible data that can be changed with the help of admin panel and user’s account. Documents search index is not included in settings – it will be updated with the help of Hes.Services.ConnectorManager.

4.2 The first setup steps

In the initial setup, the following steps are recommended:

1. Connect to the domain – authorization context (see 4.6.1). You must enter these settings before further steps will be executed!
2. Register connectors (see 4.3).
3. Add storages of documents (see 4.4).

4.3 Connector registration

When you first sign in no connectors are available in the list. HES package includes several types of connectors: for the connection to the file system of the company, to the web resources, etc. There is the possibility to add your own custom connectors using the HES Connector API (see 6 – Useful links).

To register the connectors, follow these steps:

1. Enter the address of the connector in the admin area on the “Connectors” page. The address of standard connectors can be found in the description of the corresponding connector (5.2 - 5.4). The port on which the custom connector are registered must be requested from the developer.
2. Click the “Register” button.

It is not necessary to register several connectors of the same type in the case of standard connectors from the package.

Notice! Check that the connector service was launched during the connector registration.

4.4 Adding a storage of documents

Once the necessary connectors have been added to the HES software, the storages, where the documents are stored, need to be added. The same connector can be connected to a plurality of storages of the respective type (multiple network storages or several local sites with the documents). In addition, the documents with the same URI can only be in one connector storage.

To add storage, select “Details” button in “Connectors” page.

Connectors

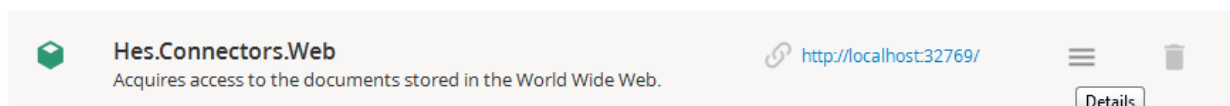


Fig. 10. Connector area, “Details”.

Next, select “Settings” tab, where operations such as adding, deleting, configuring of storages can be performed. When you add new storages, you need to specify their settings. Details of connectors and their settings are described in sections 5.2 - 5.4.

4.5 Administration Area

Pages of admin panel contain navigation area and workspace. The workspace contains elements for options edit and display of system operation information.

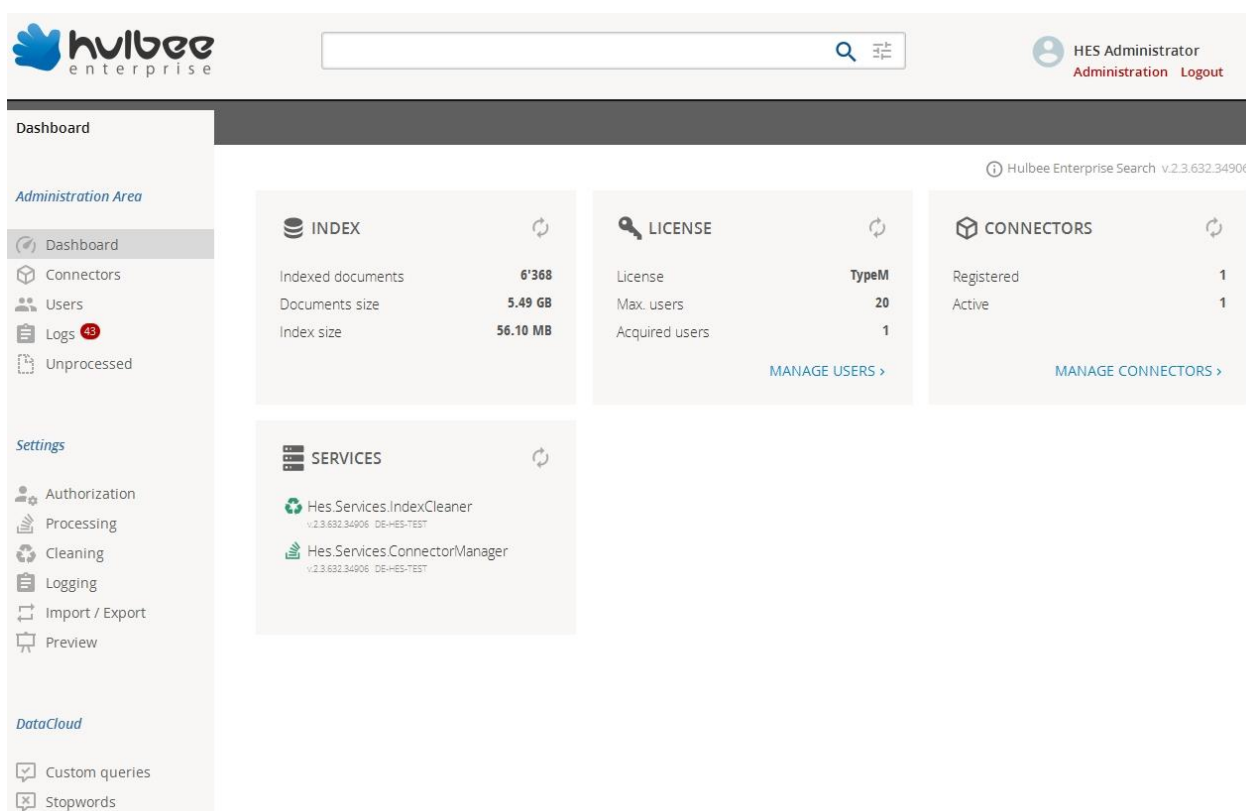


Fig. 11. Administration area. General view.

Here are the screenshots with areas, which change when you follow different links of the admin panel.

4.5.1 Dashboard

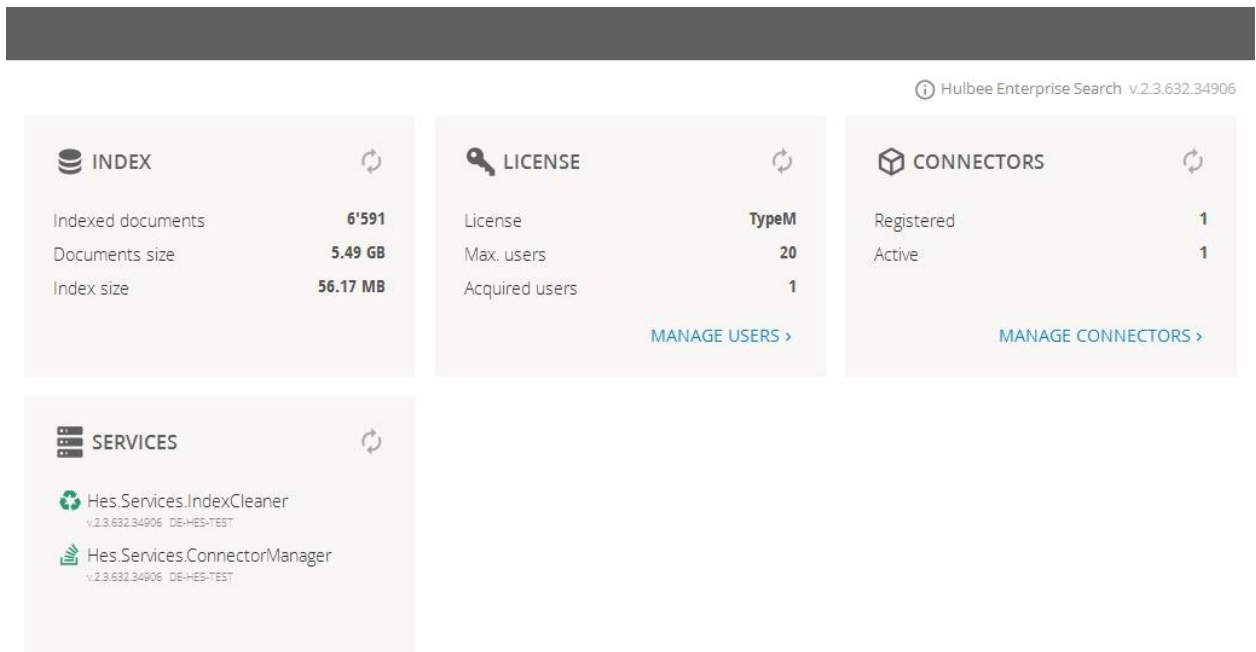


Fig. 12. "Dashboard" page.

The page shows widgets on which you can see different kinds of statistics.

"INDEX" widget shows information about an index, namely:

- **Indexed documents.** The number of indexed documents. It can differ from the total number of files on all of the available resources, as a part of documents does not need to be processed because of different filters (by size, by extension). At the same time, archives and mail messages may amount to more than one file.
- **Documents size.** The overall size of documents in the index. The sum of their binary sizes is meant. It can differ from the space that the all of the available resources occupy by the same reasons as for preceding item.
- **Index size.** The size of Elasticsearch index on the disk.

"LICENSE" widget shows license type, the number of users who are already using the service and the number of users who potentially can begin using it.

"CONNECTORS". Information about connectors that are registered in the system.

"SERVICES". Provide an opportunity to check if Connector Manager and Index Cleaner are functioning.

4.5.2 Connectors

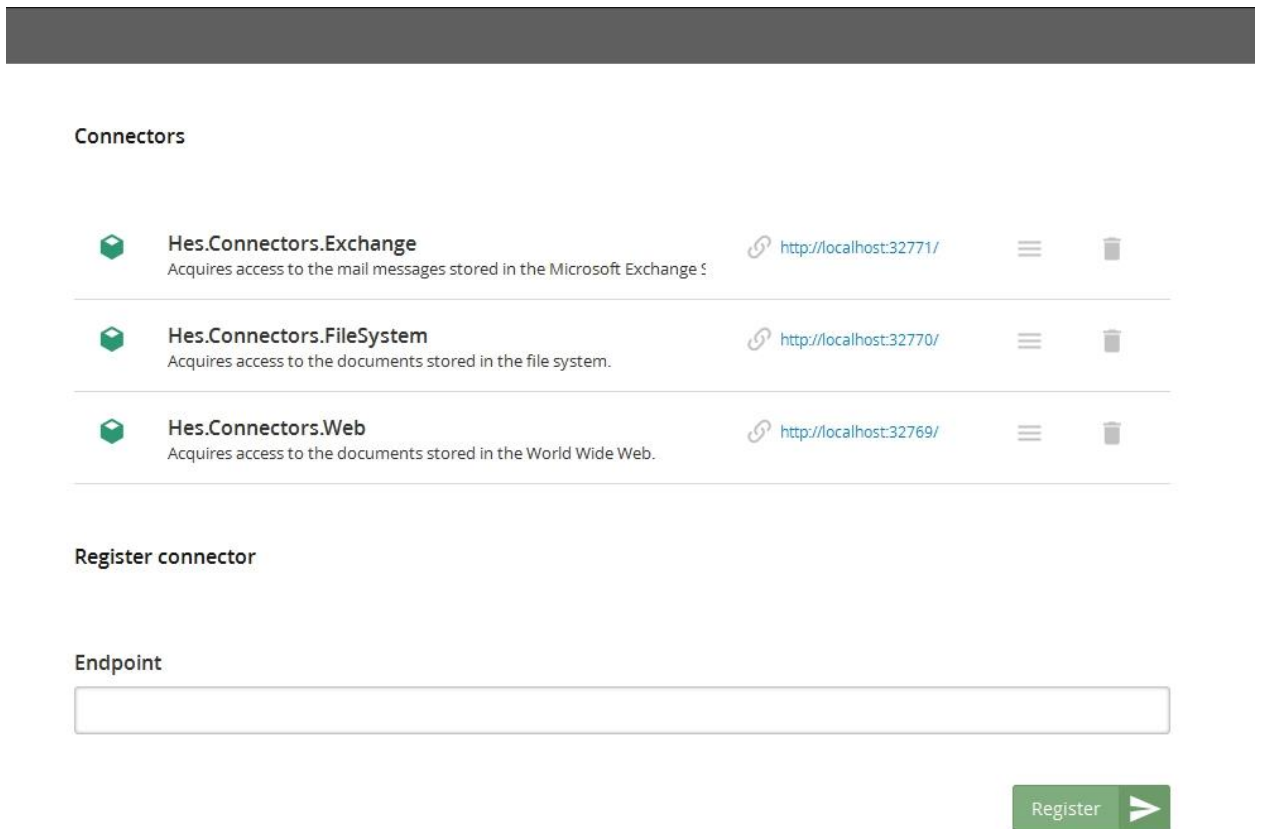


Fig. 13. "Connectors" page.

Connectors page shows information about connectors that are registered in the system.

It also contains the "Details" button, where you can access the settings of the corresponding connector. The same settings are available after clicking on the connector.

By default, working with connectors is carried out by http protocol. The field "**Endpoint**" is intended for input of the URL of the specific connector which is taken from the configuration file in the case of the usage the connectors from the HES package. The URL of a custom connector must be requested from the developer of the connector.

Existing connectors can be unregistered (see. 5.1.4) or new ones registered (see 4.3).

4.5.3 Users

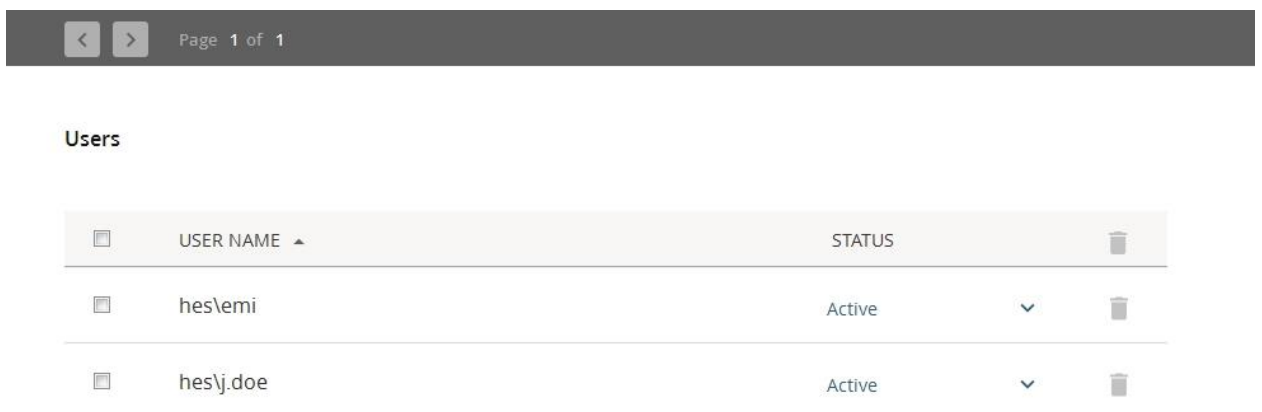


Fig. 14. "Users" page.

Users page allows to see authorized users. Here it is also possible to forbid some users to conduct a search in the HES system. In this case, this user will not be counted during verification of the compliance of the number of users with the license conditions.

4.5.4 Logs

Page 1 of 5			
Critical, Error, Warning, Information...			
Hes.Services.ConnectorManager, He...			
TIMESTAMP	EVENT	MESSAGE	PROCESS
Today, 11:32:27	Stop	Finished processing of the feed `data` (Hes.Connectors.FileSystem).	Hes.Services.ConnectorManager
Today, 11:30:44	Stop	Finished processing of the feed `store` (Hes.Connectors.FileSystem).	Hes.Services.ConnectorManager
Today, 11:30:07	Start	Started processing of the feed `data` (Hes.Connectors.FileSystem).	Hes.Services.ConnectorManager
Today, 11:29:06	Start	Started processing of the feed `store` (Hes.Connectors.FileSystem).	Hes.Services.ConnectorManager
Today, 11:20:00	Information	Scheduled cleanup of the source `Hes.Connectors.FileSystem:sh`.	Hes.Services.IndexCleaner
Today, 11:19:20	Warning	The feed sh processing is suspended or not configured.	Hes.Services.ConnectorManager

Fig. 15. "Logs" page.

This page shows error messages, warning messages and other information. You may filter event type or services by ticking appropriate items in the drop-down lists Events/Processes respectively.

4.5.5 Unprocessed

Page 1 of 2			
Unprocessed Documents			
The list of documents that have not been processed due to some errors. You can trigger some items to try process them once again or schedule processing with extended limits for the extremely important documents.			
<input type="checkbox"/>	STATUS	URI	
<input type="checkbox"/>	Aborted	\\172.16.16.242\data\Swisscows.1.14.628.34547.zip File is to large to be processed.	<input type="button" value="Repeat once"/>
<input type="checkbox"/>	Aborted	\\172.16.16.242\data\Swisscows.1.14.629.34558.zip File is to large to be processed.	<input type="button" value="Repeat once"/>

Fig. 16. "Unprocessed" page.

The list of documents that have not been processed due to some errors.


File processing is limited to “Watchdog timeout” (see 4.6.2). If because of the timeout the document is placed into the list of unprocessed files, you can schedule processing with extended limits (maximum – 30 minutes) for the extremely important documents.

Also on “Unprocessed” page you can trigger some items to try process them once again.

4.6 Settings

4.6.1 Authorization

Authorization context
The type of store for the principal context (server or domain) against which all requests are performed.

Domain (Active Directory Domain Services store) 

Domain name
The name of the domain or server.

hestest.local

NetBIOS Domain Name
NetBIOS domain name (by default, the leftmost label in the DNS domain name up to the first 15 bytes)

hestest

Container
The container on the store to use as the root of the context.

CN=Users,DC=YourCompany,DC=com

Username
The username used to connect to the store.

hes

Password

The password used to connect to the store.





Access Rules

Grant or deny access to HES application for certain users and groups.

	IDENTITY	ACCESS	
+	User or group SAMAccountName, e.g. DOMAIN\Everyone	Allow	▼

API Tokens

Grant access to HES Search API using OAuth 2 access tokens

	CLIENT ID	CLIENT SECRET	
	Default_API_Client	cdd9d834f5e153788c86ee9d92b6b379	
+	Type new Client ID		

Save Settings



Fig. 17. "Authorization" page.

Here data should be entered for the connection of HES to Active Directory. Next, this server will be used for the extraction of the user data and permissions for accessing documents.

The "Domain" option must be selected as authorization context. **"Authorization context"** encapsulates the server or domain against which all operations are performed. It is the container that is used as the base of those operations, and the credentials used to perform the operations.

Also, the "Machine" option as a test mode is available. When selecting the "Machine" authorization, demand is placed on the local database of user accounts on the computer, on which HES is located. This can be used in smaller networks that do not use Active Directory, or for test purposes. Usually, it is not used in conventional scenarios.

If the machine, on which the Processing server is installed, is not entered in the Active Directory domain, there may be problems in determining the domain controller IP-address (since in the settings it is represented in the symbolic form). To solve this problem, please proceed as follows:

1. Go to the Control Panel.
2. Click Network and Internet > Network and Sharing Center > Change adapter settings.
3. Select the connection for which you want to configure DNS.
4. Select the Networking tab. Under this connection uses the following items, select Internet Protocol Version 4 (TCP/IPv4) and then click Properties.
5. Click Advanced and select the DNS tab.

6. As “Preferred DNS server” add the IP address of the server on which the DNS service of the destination domain is provided.

Next, the authorization fields should be filled in:

- **“Domain name”**: Domain name of the directory service.
- **“NetBIOS Domain Name”**: NetBIOS name of the directory service.
- **“Container”**: The path to the container on the store with the user from Active Directory, whose data will be specified in the fields “Username” and “Password”. This field is optional. The field may be not empty when it is necessary to restrict access to certain groups/users subset, or if there are multiple catalogs in the overall structure of Active Directory. Example of the path to the container: “CN=Users,DC=hes,DC=hulbee,DC=com”.
- **“Access Rules”**: In this area the users (groups) to which access to HES is allowed or denied can be accurately determined. If the field remains empty, the access applies for all users of the domain.
- **“API Tokens”**: Grant access to HES Search API using OAuth 2 access tokens. This API allows the user to provide an alternative interface to HES search (for example, to insert it into various CMS). Setting contains Client ID – string value to specify a specific API, and Client Secret – the value that is automatically generated for each specific client ID. The obtained pair “Client Secret:Client ID” is used for generation of API token and passed on by the administrator to developers who need it to work with the applications. An application that uses a specific API token will have access to the data that the specific user (username and password) has access. It must be an existing user of the domain. HES Administrator cannot act as such user. The documentation for the HES Search API is available at: {URL_to_HES}/api/docs.

In the fields **“Username”** and **“Password”** not only the data of the administrator can be entered, but also any user from the Active Directory that has the appropriate access rights.

4.6.2 Processing

Concurrent workers count

The maximum number of concurrent operations that can be run by the processing engine.

2

^

v

Watchdog timeout

Time to wait for the results from worker before the processing task will be canceled. The value must be in format "HH:MM:SS", e.g. 00:05:00 - five minutes.

00:05:15

Idle timeout

Time to wait for the tasks from the Connector Manager before the worker will be shutted down. The value must be in format "HH:MM:SS", e.g. 00:05:00 - five minutes.

00:05:15

Fig. 18. “Processing” page.

Select “Processing” area and tune following fields:

All rights reserved. © 2015-2017 Hulbee AG.

24/46

- **“Concurrent workers count”**: The maximum number of concurrent operations that can be run by the processing engine. To get better processing speed, you can enter here a number up to the CPU cores number. For configuration with Elasticsearch, located on another server, it is recommended to assign the number of workers in accordance with the number of cores in the machine with Elasticsearch. But this number should not be greater than the number of cores in the Processing server. For configuration of HESbox and in the cases when all modules are located on one machine, it is recommended to set the number of workers up to the half of the CPU cores number.
- **“Watchdog timeout”**: Time to wait for the results from worker before the processing task will be canceled. Set the time base value (it should be enough to process most document repositories, e.g. 15 ... 30 seconds). It corresponds to the duration of the first iteration of the processing of the document. It can be quite a little time for the first indexing cycle to process all documents quickly. If a document could not be processed in the specified time, there is a forced break and the next document is processed. In the next round, the time is automatically increased by three times for documents whose processing was terminated by a timeout. On the third try – twelve times. This allows to process heavy files (archives, large documents), which have not been processed during the first two cycles. Further attempts are not made. The unprocessed documents are shown on the page “Unprocessed” (see 4.5.5).
- **“Idle timeout”**: Time to wait for the tasks from the Connector Manager before the worker will be shut down.

Max. document size

The documents whose size in bytes exceeds the given value will not be processed.

Max. extracting content size

The maximum size of content in bytes that can be extracted from document.

Max. attachments size
















The maximum size of document attachments in bytes.

Fig. 19. “Processing” page. Continuation.

- **“Max. document size”**: The documents whose size in bytes exceeds the set value will not be processed. Such documents are added to the list of unprocessed files (see 4.5.5), but their metadata are extracted.
- **“Max. extracting content size”**: Maximum size of text in symbols, which undergoes further processing and indexing. Text extracted from document is meant and not binary file size. Remaining text is cut.
- **“Max. attachments size”**: Maximum size of attached files. If the size of the attached file or archived document exceeds the predetermined size, it is not processed but its metadata are extracted.

Extraction method by extension

Content extraction method depending on document extension.

	EXTENSION	FILTER	
	pdf	IFilter then native	
	docx;doc;docm	IFilter then native	
	ppt; pptx	IFilter then native	
	xls; xlsx	IFilter then native	
	one	IFilter then native	
	csv	Native only	
+	<input type="text"/>	IFilter then native	

Processing limits by extension

Limits on the documents size depending on extension.

	EXTENSION	LIMIT	
	avi;wmv;mp4;mkv	8589934592	
+	<input type="text"/>	107374182400	

Save Settings 

Fig. 20. "Processing" page. Finish.

- **"Extraction method by extension"**: Selection of converter type depending on extension. As a rule, IFilter is slower, but it is valid for office documents and pdf documents.
- **"Processing limits by extension"**: Limits on the documents size depending on extension. The files of some types are large and fall within size limitation. But their processing is quite easy, as a little part of data is processed. It mainly refers to media files "avi;wmv;mp4;mkv". For this purpose, you may set individual limitations for this group of files. You may enter them in one line, separating extensions by ";" symbol.

After changing these options, click "Save Settings" button.

Notice! Before adding storage, settings for "Processing" should be customized, as usage of some of them may require a re-indexing of storage.

4.6.3 Cleaning

Concurrent workers count

The maximum number of concurrent operations that can be run by the cleaner. It is recommended that this value wasn't less than the number of registered connectors.

Connectors discovery interval

The interval between connectors discovery sessions. The value must be in format "HH:MM:SS", e.g. 00:05:00 - five minutes.

Save Settings

✓

Fig. 21. "Cleaning" page.

Select "Cleaning" area and adjust following fields:

- **"Concurrent workers count"**: The maximum number of concurrent operations can be run by the cleaner. It is recommended that this value was not less than the number of registered connectors.
- **"Connectors discovery interval"**: The interval between sessions, during which the cleaner checks which connectors are available. It is recommended to perform this test not too often (no more frequently than once every 5 minutes), because while being performed a query to the index is made. This procedure may slow down search conducted by users at the time. The value must be in format "HH:MM:SS", e.g. 00:05:00 - five minutes.

After changing these options, click "Save Settings" button.

4.6.4 Logging

Trace level

Information

Automatically remove logs older than

Week

Save Settings ✓

Event types

Critical, Error, Warning, Information, Verbose, Start, Stop

Before date

11.07.2016

Delete

Fig. 22. “Logging” page.

You can change the following parameters on the Logging page.

- **“Trace level”**: Drop-down list, where we can indicate the events that should be recorded in the diagnostics. Hes.Services.ConnectorManager and IndexCleaner services will use new settings after restart.
- **“Automatically remove logs older than”**: This option allows you to indicate, how long HES logs should be stored. Both records stored in the Indexstore (they are shown in the admin panel) and stored as text files on the machine with the Processing server (typical place – C:\ProgramData\Hulbee AG\HES\Logs) are removed. Over time they can occupy too much space, that is why it is not recommended to turn this option off (save in exceptional circumstances). IndexCleaner service removes logs. If you want to change a retention period, you need to restart this service.

You can also remove some events (**“Event types”**), kept in Indexstore, manually. In order to get that done, you need to choose types of events that should be removed and date (**“Before date”**) to which it is necessary to remove them. Then click **“Delete”** button.

4.6.5 Import / Export

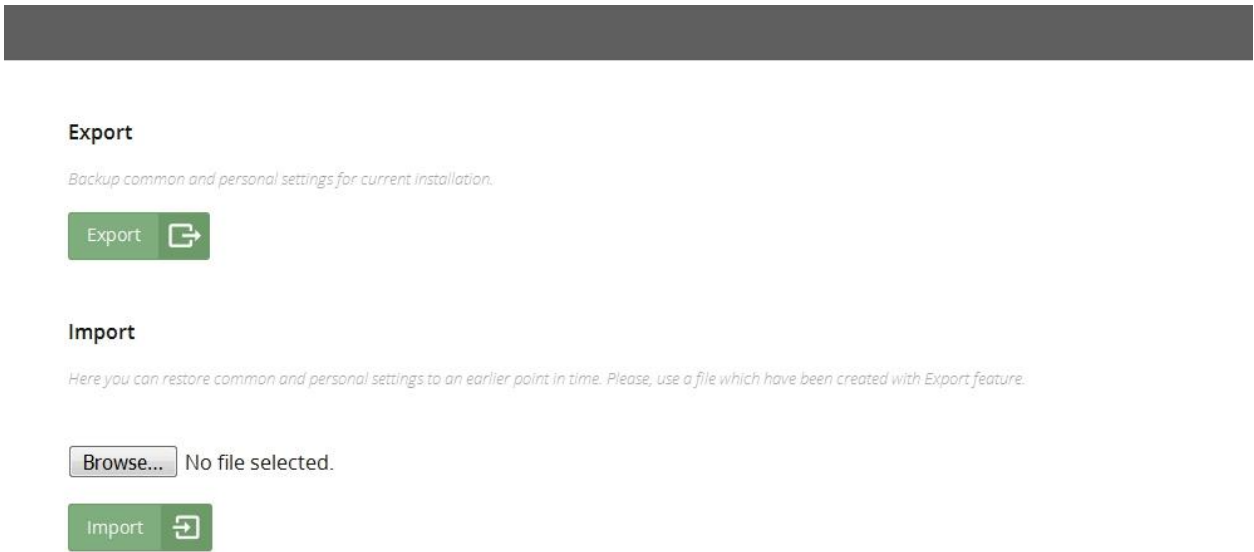


Fig. 23. “Import / Export” page.

This page contains commands, which provide an opportunity to save HES settings into a file and restore them when necessary. Both common and personal settings are saved. These commands will be useful in the case of system restore after failure, or after update installation.

When importing settings, connectors must be registered manually.

4.6.6 Preview

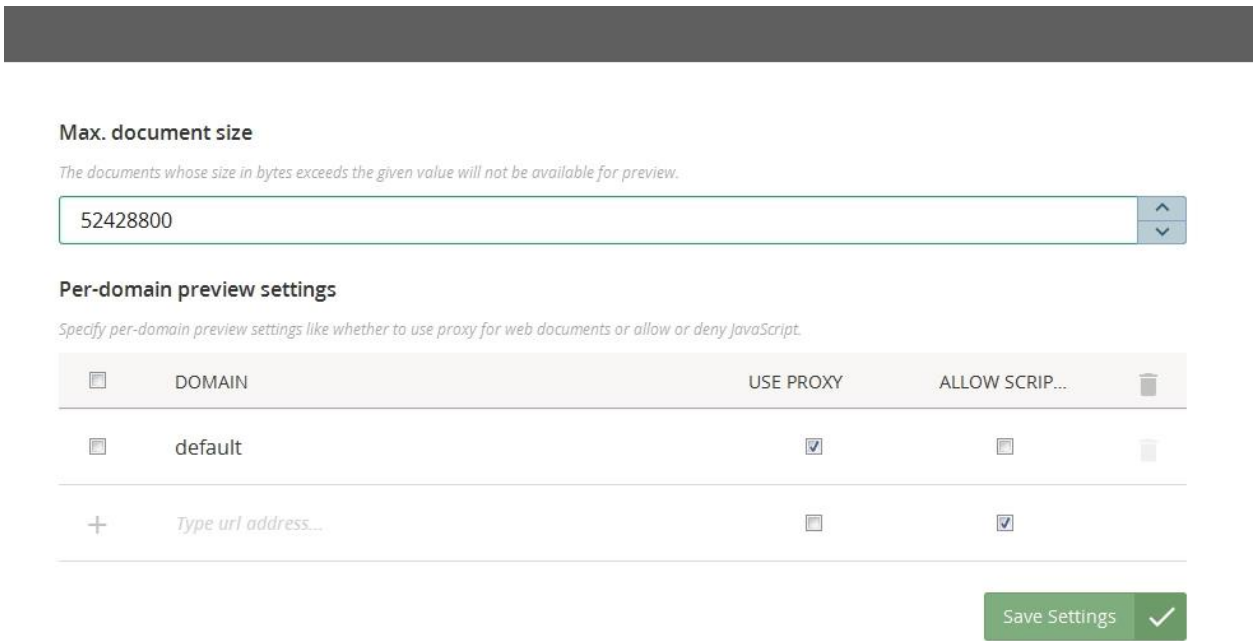


Fig. 24. “Preview” page.

This page allows you to enter the necessary settings for correctly displaying the preview of some types of documents (HTML, PDF, MS Word, etc.).

In “Preview” section, the following parameters can be set:

- **“Max. document size”:** The maximum size of the document, which is still available for the preview mode. If the size of the document exceeds the predetermined size, the tab “Preview” will be provided with appropriate notification instead of the document. The limitation applies to all file formats, but it makes sense for large text documents that are uploaded as entire file.
- **“Per-domain preview settings”:** In DOMAIN field enter the base address (base URL) (protocol, domain, and optional path) for the pages to which these settings should apply. If the page does not match any one of these base URL, the rule is applied, indicated by the label “default”.

ALLOW SCRIPTS Column is intended to prohibit or permit to run scripts and is only applicable for the html-document. Allow scripts only desirable for its proven resources to avoid various kinds of attacks.

The files which are available to the user on the protocol file:// (on a local network file storage), in any case, will be shown through the Proxy, because the user's default browser is forbidden to work with the file system for security reasons.

Proxy must be used in the following cases:

- The resources are located in the file storage (file:// protocol).
- If HES interface works via the https-protocol, and all (or part) of resources to be displayed in “Preview” tab, are available through the http:// or file:// protocol. Since the preview rendering takes place in the context of the user's browser, such resources are not available for security reasons (as in the previous item).

In other cases (e.g., if HES is available via the http-protocol or it is not necessary to display documents from the file storage), it makes sense to work directly, without Proxy. Among other things, this makes it possible to preview web pages from the websites with authorization.

To add Proxy settings for your website follow these steps:

1. Enter your website URL into the “type url address...”
2. Take out the check mark in the USE PROXY column.
3. Save settings.

If before viewing the page of the site on the “Preview” tab you are prompted for an authentication, type the username and password.

4.7 Datacloud

These settings are similar to settings in the user account, but they concern all HES users (instead of one).

4.7.1 Custom queries












  Page 1 of 1			Show: 20 
Custom queries			
	CUSTOM QUERY ▲		
	custom_query1		
	custom_query2		
	Type new Custom query...		

Fig. 25. “Custom queries” page.

“Custom queries” option allows to add keywords, later on appeared in DataCloud on home pages of all users. Keywords help to enter typical search queries quickly.

You may add, edit and delete custom queries. The custom query navigation is available.

4.7.2 Stopwords












  Page 1 of 140			Show: 20 
Stopwords			
	STOPWORD ▲		
	2nd		
	about		
	Type new Stopword...		

Fig. 26. “Stopwords” page.

Stopwords imply a list of words that are often found in processed documents but do not carry any additional information to the search engine. Their presence may have a negative impact on the relevance of the request.

Using “Stopwords” tab, you can indicate words, which will not participate in the further processing of the request. They also will not appear in the DataCloud, located on the search results page. For example, the name of the user’s company. It can be found almost in every document, for which reason it is useless for query specification.

Stopwords, added by the administrator, will be cut out from the request only after updating the index structure (triggered by pressing the button “Synchronize”). During synchronization, no search can be performed. HES user cannot carry out this setting.

Work with Stopwords list is similar to the work with search query list in the “Custom queries” tab.

5 Connectors

5.1 Common tunes

Connectors in HES are running as a Windows service that are installed using the Network Setup Wizard. They are by default installed in the following folder C:\Program Files\Hulbee AG\Hulbee Enterprise Search\Connectors\... .

Connectors can be configured in two ways. Most settings are available via the HES admin area. The respective part of the configuration (such as integrated port, authentication, etc.), is inserted in their configuration file, under the name <Name of the executable file> .config.

5.1.1 Settings of the connectors via the configuration files.

The basic configuration options that can be useful in the settings:

<add key="Service.EndPoint" value="http://+:32770" /> – the protocol and the port, which are used in the interface for the realization of the HES Connector API.

The connector can operate over http and https protocol. It may perhaps be necessary to reserve the existing protocol and port by using this command (starts in administrator mode):

```
netsh http add urlacl url=http://+:32769/ user=\Everyone
```

HES Installer executes this operation automatically for http protocol. Run it manually if you want to change the port or protocol of Connector.

Other useful commands:

```
netsh http add urlacl url=https://+:32769/ user=\Everyone - similarly for  
https protocol  
netsh http show urlacl - show URL that are reserved  
netsh http delete urlacl url=http://+:32769/ - delete redundancy of the  
specified address
```

At the same time for the https protocol, you must install the SSL-certificate. This can be a full certificate, which the company bought from any authorized distributor. You can also use a self-signed certificate. To create a self-signed certificate you need to follow these steps:

1. Creating a certificate. Enter in the command line of the PowerShell, running with Administrator privileges:

```
New-SelfSignedCertificate -DnsName localhost -CertStoreLocation  
Cert:\LocalMachine\My
```

In response, the hash of the certificate is displayed. Example:

3214979BE7BD608A426404537FCDB90103E157DB

2. Converting the certificate into the trusted status. In order for a certificate to be trusted, install it in Cert:\Local Machine\Root. Run the commands in PowerShell:


```
$cert = (get-item Cert:\LocalMachine\My\*)
$store = (get-item Cert:\LocalMachine\Root\ )
$store.Open("ReadWrite")
$store.Add($cert)
$store.Close()
```

where * – hash of the certificate that you created in 1. Step.

3. Installing the certificate to the same port that is running the connector. To install a trusted certificate in PowerShell, run the following command:

```
netsh http add sslcert ipport=0.0.0.0:32769 certhash=* appid='**'
```

where * – hash of the certificate that you created in 1. Step, ** – any valid GUID (for its generation can be used, for example, <https://www.guidgenerator.com/>).

Authentication can be configured using “authentication” area (the corresponding data must be entered in the connector settings on the “GENERAL” tab):

```
<authentication>
  <!-- Basic authentication section:

  <basic enabled="true or false" username="allowed username" password="allowed
    user password" />

  The following code example demonstrates how to allow access to user with
    name "foo" and password "bar".

  <basic enabled="true" username="foo" password="bar" />
  -->
  <basic enabled="true" username="admin" password="pass"/>

  <!-- Windows authentication section:

  <windows enabled="true or false">
  <allow users="comma-separated list of users" roles="comma-separated list of
    roles" />
  <deny users="comma-separated list of users" roles="comma-separated list of
    roles" />
  </windows>

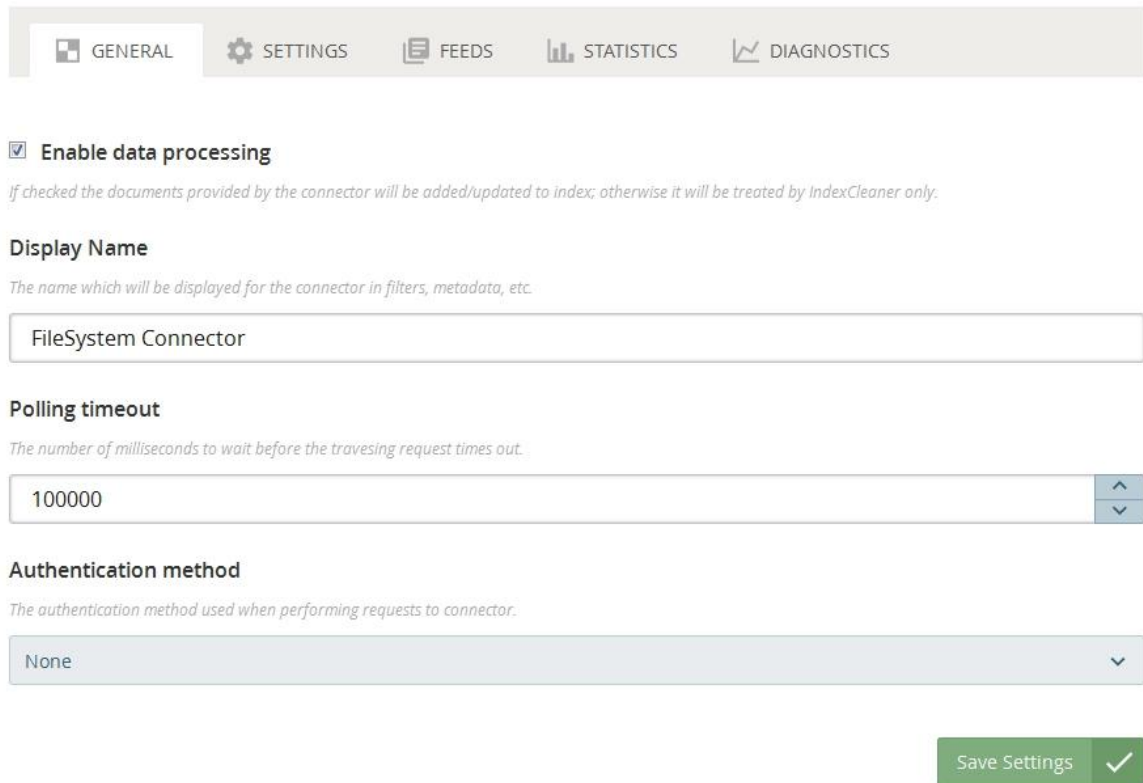
  The following code example demonstrates how to allow access to all members
    of the Admins role
  and deny access to all other user accounts.

  <windows enabled="true">
  <allow users="DOMAIN\Administrators" />
  <deny users="*" />
  </windows>
  -->
  <windows enabled="false" />
</authentication>
```

5.1.2 Settings of the connectors in the admin area.

In the admin area under “Connector” there is the “Details” button (Fig. 10), which allows you to go to the tabs – the settings for connectors.

There are the following settings on the “GENERAL” tab:



The screenshot shows the 'GENERAL' tab of a connector settings page. At the top, there is a navigation bar with five tabs: 'GENERAL' (selected), 'SETTINGS', 'FEEDS', 'STATISTICS', and 'DIAGNOSTICS'. Below the navigation bar, the 'Enable data processing' option is checked, with a note: 'If checked the documents provided by the connector will be added/updated to index; otherwise it will be treated by IndexCleaner only.' The 'Display Name' field is set to 'FileSystem Connector', with a note: 'The name which will be displayed for the connector in filters, metadata, etc.' The 'Polling timeout' field is set to '100000', with a note: 'The number of milliseconds to wait before the traversing request times out.' The 'Authentication method' dropdown is set to 'None', with a note: 'The authentication method used when performing requests to connector.' At the bottom right, there is a green 'Save Settings' button with a checkmark icon.

Fig. 27 Connectors. “GENERAL” tab.

- **“Enable data processing”**: Whether data coming from connector should be processed. If checked the documents in the index will be indexed, updated, added; otherwise, documents are only checked for removal from the index by the IndexCleaner, but an active traversal of all storages of the connector will not be made. The same setting is available in the “FEEDS” tab (Fig. 28) for each storage.
- **“Display Name”**: Name of the connector, which is set by the administrator. It is displayed on the user’s page in the filter area and on the page of the advanced search.
- **“Polling timeout”** (milliseconds): The waiting time for the response from the connector. For some connectors, which take a long time for the data obtaining, it may be that the time limit for handling the batch of data is insufficient. If this time is exceeded, the storage will be re-crawled and the search engine will display an error message in the “Logs” section. To continue the crawling on the next bypass, the polling timeout must be increased.
- **“Authentication method”**: The authentication method used when performing request to connector. If the connector is configured in the authentication mode (set in the configuration file of the connector - see section 5.1.1), here the authentication settings must be set so that HES core has access to the connector.

In the “SETTINGS” tab storage settings are displayed. They are different for specific connectors from the HES package and are described in the following sections: 5.2.1, 5.3.1 and 5.4.1.

The list of settings on the “FEEDS” tab is available as soon as they are opened in one of the storage lists (Fig. 28).

GENERAL

SETTINGS

FEEDS

STATISTICS

DIAGNOSTICS

store

http://localhost:32770/feed/store

Display Name
The name which will be displayed for the feed in filters, metadata, etc.

store

TRAVERSING

☒ **Enable traversing**
If checked the documents provided by the feed will be added/updated to index.

Polling batch size
The count of documents to fetch per traversing request.

10

Polling interval
Minimal timeout between traversing requests in milliseconds.

1000

Polling session timeout
Minimal timeout between traversing sessions in seconds.

900

Web requests interval
Minimal timeout between web requests to obtain http(s)/ftp contents in milliseconds.

0

CLEANING

☒ **Enable cleaning outdated data**
If checked the cleaner service will periodically revise indexed documents and remove obsolete data.

Cleaning batch size
The count of documents to analyze per request.

1000

Polling interval
Minimal timeout between cleaning requests in milliseconds.

10000

Cleaning request timeout
The number of seconds to wait before the cleaning request times out.

100

Save Settings ✓

Fig. 28 Connectors. “FEEDS” tab.

- **“Display Name”**: The name of the feed, which will be displayed in filters area, metadata, etc.

“TRAVERSING” section of the “FEEDS” tab:

- **“Enable traversing”**: Whether the documents provided by the feed will be added/updated to index.
- **“Polling batch size”**: The count of documents to fetch per traversing request.
- **“Polling interval”** (milliseconds): Minimal timeout between traversing requests.
- **“Polling session timeout”** (seconds): The minimal time interval after a full pass of a storage until a new round can start. If you know a data in storage are rarely updated, it makes sense to increase session timeout. This way, the number of simultaneously occurring bypasses of the engine is reduced (decreases the load on the engine). The default value is 15min.
- **“Web requests interval”** (milliseconds): Minimal timeout between web requests to obtain http(s)/ftp contents. With the help of this setting, you can regulate a site loading (set timeout more, making the loading less and vice versa).

As a recommended value Polling interval, it proposed a value that is not less than (Polling batch size * Web requests interval).

“CLEANING” section of the “FEEDS” tab:

- **“Enable cleaning outdated data”**: Whether the cleaner service will periodically revise indexed documents and remove obsolete data.
- **“Cleaning batch size”**: An approximate number of documents to analyze per request for the removal of obsolete data from the index (cleaning request). The real value may vary depending on the number of documents in the feed. Recommended value for this setting is 1000.

- **“Polling interval”** (milliseconds): Minimal timeout between cleaning requests.
- **“Cleaning request timeout”** (seconds): The maximum waiting time for a response from the connector after cleaning request. For connectors that take a long time to respond to cleaning requests, the time can be increased. Recommended value for this setting is not more than 100 sec.

The “STATISTICS” tab contains a combination of “name” - “value”. The set of statistics data is defined by the developer of the connector.

The “DIAGNOSTICS” tab displays last error messages that are happened in the connector.

5.1.3 Manual connector setup

Connectors represent Windows services that are also REST services. By default, all standard connectors are installed and run with HES core. If for some reason you want to move the connectors to another machine or install another copy of the connector on the current machine, you can do this as follows:

1. Copy the folder with the connector to the desired location.
2. Reserve under the connector free port and protocol (as described in section 5.1.1).
3. Set it in the configuration file of the connector.
4. Install the connector as a Windows service on behalf of the Network Service user using an unused service name. Installation should be performed by the executable file of a connector with additional parameters. The syntax of additional parameters can be obtained by running the service with “help” parameter.

The procedure described above is only for standard connectors that come with the HES packaging. Connectors developed by third-party developers are installed and configured according to the instructions that come with these connectors.

5.1.4 Connector unregistration

If a connector is not used, it can be removed from the list of registered. This is done by clicking on the “bin” icon next to the appropriate connector (section “Connectors” of the administrative panel).

Connectors

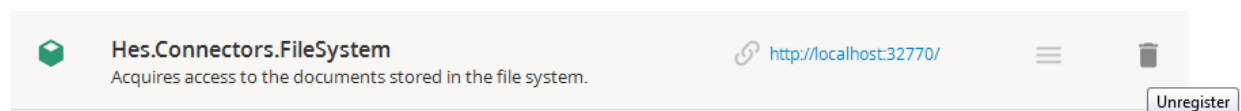


Fig. 29. Connector area, “Unregister”.

After unregistering the connector, all the documents that relate to it are removed gradually from the index.

However, it must be considered that an unregistered connector continues to function as a service. This may cause some connector types (Web Connector) to continue to bypass the resource automatically. To prevent useless load, either before you unregister the connector, remove the storages configuration from it, or simply stop (and disable automatic start on reboot) the corresponding Windows Service. In the latter case, one has to consider that before the re-registration of the connector it must be run again.

When deleting a connector from the list of registered and after reactivation, the storage settings in the “SETTINGS” tab will be saved, but the settings for bypassing storages (tab “FEEDS”) return to the default settings.

5.2 FileSystem Connector

This is used to access documents that are located on the local file system. If installation by default is used, FileSystem Connector is available at <http://localhost:32770/>.

5.2.1 File connector settings.

The settings for the FileSystem Connector include:

- **“Storage name”**: Name of the storage, which should be processed. If it is changed, the storage is run again.
- **“Scan location”**: Path on the local machine or in the local area network, which should be scanned.
- **“Default checkpoint”**: The path to the directory which is specified in the “Scan location”, is the starting point for going through the storage. If the value is not specified, the storage is scanned from the beginning. The “Default checkpoint” setting makes sense if an error occurs during the crawling of the storage for some reason and the crawling will start again.
- **“Scan/watch files wildcard”**: The search pattern to match against the names of files in scan location. Only documents that match the specified patterns are indexed. You can enter only extension or only the name (part of the name) to search the file. Example: ***.tmp** or **doc*.* (doc*)** or **document.***. Every single wildcard pattern is written into a separate line.
- **“Ignore patterns”**: The pattern to skip files or directories in scan location. (Leave empty to do not skip any file or directory). Example: ***.tmp** – ignore files with a specified extension or ***/secure_folder/*** – ignore files whose paths this folder contains. Every single Ignore pattern is written into a separate line.
- **“Access credentials”**: The identity credentials, which will be used for access to storage. User access rights are defined according to the security permissions on files and folders in the file system. If the connector cannot connect to the repository, a message will be recorded in the diagnostic of the connector.
- **“Domain”, “Username”, “Password”**: Data of the user, who is defined in “Access credentials” and has access to the repository, which specified in “Scan location”.
- **“Feed traversing timeout (in milliseconds)”**: The maximum time that the connector can take to collect a batch of documents to be processed. If the timeout has been exceeded, the number of documents collected by then is displayed. This collection optimizes network load and makes sense in the case of a large number of empty folders and folders with 1-2 files in the file store. It is recommended to set a value several times smaller than the typical network timeouts (5-10 seconds). A value of 0 means that the answer is formed at once, regardless of the number of files in the directory (but not more than the value “Polling batch size”).
- **“Maximum count of files in directory”**: Maximum count of files in directory (not including subdirectories) for processing. If the number of files in the directory is larger, only what is specified is indexed, and a warning is displayed in the “DIAGNOSTICS” tab. If some large directories still need to be scanned, this value should be increased, as well as the value in the “Polling timeout” field in the “GENERAL” tab. If zero, this setting has no effect.

5.3 MS-Exchange Connector

Allows search engine to index the content of emails of MS-Exchange. In the section for the connector, no separate storage facilities exist, as any instance of MS-Exchange allows you to work with mailboxes of any domain user, even if they are physically located in another MS-Exchange Server.

Requirements to Exchange service for connector: Remote PowerShell must be available on it (so far only basic authorization is supported).

If installation by default is used, Exchange Connector is available at <http://localhost:32771/>.

5.3.1 MS-Exchange Connector settings.

The settings for the MS-Exchange Connector include:

- **“Exchange Server PowerShell url”**: URL of Exchange Server PowerShell. Example: <https://my-exchange.com/PowerShell>.
- **“Use internal link for OWA”**: Whether to use an internal URL OWA (Outlook Web App), while saving an URL to the document. These links will be stored in the index and used to open documents in the MS-Exchange web-interface. When switched off, an external reference is used. External and internal URL may differ. The configuration of internal and external URL is performed through Exchange Admin Center.
- **“Use internal link for EWS”**: Whether to use the internal URL for EWS (Exchange Web Services) to enable connector to communicate with the Exchange Server.
- **“Access credentials”**: The identity credentials, which will be used to access the MS-Exchange Server.
- **“Domain”, “Username”, “Password”**: Data of the user is entered, who has read access to every MS-Exchange mailbox. This user must have its own mailbox on MS-Exchange service and impersonation rights over mailboxes. In addition, the user must be a member of the group View-Only Organization Management (or Organization Management) to access a list of other users.
- **“Containers”**: Active Directory Containers on the store that will be used for searching users' mailboxes. An example of the path to the container: “CN=Users,DC=hes,DC=hulbee,DC=com”. You can simultaneously specify multiple such paths, placing each into a separate line. If the list is empty, the mailboxes of all domain users are indexed.
- **“Request timeout”** (milliseconds): The waiting time from MS-Exchange Server for a traversing request. If this time is exceeded, a bypass of the storage is started again, and the search engine will display an error message in the “Logs” section. To continue the crawling on the next pass, the polling timeout must be increased.

5.4 Web Connector

Provides access to the documents stored on the World Wide Web. It contains a crawler that traverses the pages in the local and global network. Crawler has its own period of crawling and crawls continuously. Checking out the connector does not stop the crawler (see 5.1.4).

All documents that are found under this crawler have the same settings of visibility for users within the storage. Access rights for individual user/group of users are defined according to the selected settings (“Access list mode”).

If installation by default is used, Web Connector is available at <http://localhost:32769/>.

5.4.1 Settings of the Web Connector

The list of settings contains settings for the connector and the crawler. The settings for the Web Connector include:

- **“Storage name”**. Name of the storage, which should be processed. If it is changed, the storage is run again.
- **“Url”**: URL of internet start page of the resource which should be crawled. This page must contain links that lead to other pages of the site.
- **“Authentication type”**: Determines which credential is required to crawl the site. Used when the website includes authentication.
- **“Login”/“Password”**: Login and password for the selected authentication type. An account must be used that displays all pages and documents.
- **“Maximum of concurrent threads”**: The maximum number of simultaneous CPU-threads that the connector can use for processing a resource.
- **“Maximum pages to crawl”**: The maximum number of pages to crawl. When this limit is reached, the crawler stops and is further only checking the meta information at sites (documents) already found.
- **“Maximum pages to crawl per domain”**: The maximum number of pages to crawl per domain. If the resource has links to an external domain, you can put restrictions on the domain for crawling. It makes sense, because when “Crawl external pages” and “Crawl external pages links” are checked, processing of pages and documents is made not only in the domain specified on the home page. If zero, this setting has no effect.
- **“Maximum page size”** (bytes): The maximum size of page for crawling. If the page size is above this value, it will not be downloaded or processed. If zero, this setting has no effect.
- **“Crawl external pages”**: Whether pages external to the root URI should be crawled.
- **“Crawl external pages links”**: Whether pages external to the root URI should have their links crawled. This setting is useful only if the previous setting is enabled.
- **“Use canonical links”**: The possibility to crawl the page, which is available through the canonical link (see 5.4.4), instead of the page opened by the original link. One case or the other depends on the availability of rel=“canonical” (Canonical URL tag) or “og:url” (Open Graph meta-tag) in the <head> section of the page. If there is no tags support on the page, then the original pages should be crawled.
- **“Http request timeout”** (seconds): Time limit for handling the resource.
- **“Follow redirects”**: Whether the request should follow the redirection if a link with redirects exists. The link itself is not added.
- **“Maximum auto redirects”**: The maximum number of redirects that the request follows.
- **“Enable Cookies”**: Whether the cookies should be set and resent with every request when crawling through website links.
- **“Enabled SSL certificate validation”**: Whether or not to validate the server SSL certificate. If true, the default validation will be made. If false, the certificate validation is bypassed. This setting is useful to crawl sites with an invalid or expires SSL certificate. Useful in the processing of https resources.
- **“Maximum crawl depth”**: Maximum levels below root page to crawl. If the value is 0, the homepage will be crawled, but none of its links will be crawled.
- **“Maximum retries count”**: The maximum number of retries for processing the file, if the file could not be processed in the set time. If the value is 0, no retries will be made.

- **“Minimum retry delay”** (milliseconds): The minimum delay between a failed http request and the next attempt to re-access the file.
- **“Respect robots.txt”**: Whether the crawler should retrieve and respect the robots.txt file rules.
- **“Respect meta robots”**: Whether the crawler should ignore links on pages that have meta-tag of nofollow: `<meta name=“robots” content=“nofollow” />`.
- **“Respect X-Robots-Tag header”**: Whether the crawler should ignore links on pages that have an http X-Robots-Tag header of nofollow.
- **“Respect anchor rel=‘nofollow’ ”**: Whether the crawler should ignore links on pages that have rel-attribute of nofollow: ``.
- **“Ignore robots.txt if root disallowed”**: When this option is enabled, robots.txt is ignored.
- **“Robots.txt user agent”**: The user agent string to use when checking robots.txt file for specific directives.
- **“Maximum robots.txt crawl delay”** (seconds): The maximum number of seconds to respect in the robots.txt “Crawl-delay: X” directive. Enabled “Respect robots.txt” must be true for this value to be used. If zero, will use whatever the robots.txt crawl delay requests no matter how high the value is.
- **“Minimum crawl delay per domain”** (milliseconds): The number of milliseconds to wait in between http requests to the same domain. The setting is needed in order not to create too much loading on the site by crawling. If, for example, 500 milliseconds is specified, then it would mean that no more than two pages are requested from the site per second.
- **“Additional headers”**: Additional http-headers for the site are written in the format *key: value*. Every single http-header is placed into a separate line. It is necessary in rare cases.
- **“Min generations before deleting”**: The number of requests of the document before it can be deleted from crawled data.
- **“Pause after cycle”** (seconds): Delay between cycles of traversing the resource.
- **“Access list mode”**: method of determining access rights. The following options are available:
 - “Public” – allows access for all HES users.
 - “Fixed Access List” – allows access for the user/group of users specified in the “Access list allowed”, excluding forbidden in the “Access list denied”.
 - “X-HES-Users only” – selects user/group of users from response header fields “X-HES-Users-Allowed” and “X-HES-Users-Denied” only. For more details about X-HES-Users see 5.4.5. Settings “Access list allowed\denied” are ignored.
 - “Prefer X-HES-Users” – allows access for users/groups of users specified in X-HES-Users fields, otherwise (if the X-HES-Users headers do not exist) – in “Access list allowed\denied”. If the X-HES-Users headers do not exist, and “Access list allowed\denied” is not configured, the document will be available only to the HES administrator.
 - “X-HES-Users and Access List” – takes into consideration the user access rights from both lists (specified in X-HES-Users and “Access list allowed\denied”).
- **“Access list allowed”**: A list of users and groups that have access to the documents in storage. In this case, you can allow access to documents for a group of users, but deny access to specific users from this group (use for it “Access list denied”).
- **“Access list denied”**: A list of users and groups that are denied access to the resource, even if they are present in the list from “Access list allowed”. That is, in an allowed group, some users can be defined who are denied access.
- **“Active Directory credentials”**: The identity credentials, which are used to access the Active Directory to convert the symbolic names of users and groups in SID. In the fields (“Domain”,

“Container”, “Username”, “Password”) will be specified the data of a user who has read rights on the list of users from Active Directory. The username in the appropriate field should precede the domain name. This section should be filled in if “X-HES-Users” feature is used and a website returns X-HES-Users headers data containing users/groups of users in symbolic names (example: X-HES-Users-Allowed:HES\Domänen-Benutzer;HES\j.doe).

- **“Taboo rules”**: A list of regular expressions for exclusion of some web pages from crawled results. Every single regular expression is placed into a separate line. This way the crawler does not crawl unimportant parts of the site (or those where it can go to the loop, avoiding an infinite number of pages). Another scenario of the feature using is splitting one site to multiple storages. You can see examples of the regular expressions at: [https://msdn.microsoft.com/en-us/library/az24scfc\(v=vs.110\).aspx](https://msdn.microsoft.com/en-us/library/az24scfc(v=vs.110).aspx). Example: section of the website containing GET parameter actpos=2, must be placed in another storage with other access settings. To do this, enter “.*actpos=2.*” in “Taboo rules” of the primary storage, so all pages containing such parameter will be skipped on crawling. And for another storage the same “URL” must be added and its own rule “.*actpos=(?!2).*”. So pages containing parameter actpos=2 will be crawled. Please note that if the regular expression is incorrectly configured, the system overloads.

Notice! In the case of changing the settings, which can lead to a contraction of the number of the results (reducing the value in the field or “Use canonical links” is enabled), it is necessary that the Web Connector runs the storage again. To do this, you can create a new storage, but this leads to a reset of established settings. To prevent this, just change its name in the “Storage name” field and save the settings.

The fields where settings described above are possible: “Crawl depth”, “Maximum pages to crawl”, “Maximum page size”, etc.

5.4.2 Fine tuning of the Web Connector

Some Web Connector settings (common to all the storages) are available through the configuration file. They are located in the “abot” section. In the “abot” area the following attributes can be useful:

- **userAgentString** – the user agent string to use for http requests.
- **isRespectUrlNamedAnchorOrHashbangEnabled** – whether or not URL named anchors or hashbangs are considered part of the URL. If false, they will be ignored. If true, they will be considered part of the URL.
- **minAvailableMemoryRequiredInMb** – uses closest multiple of 16 to the value set. If there is not at least this much memory available before starting a crawl, throws Insufficient Memory Exception. If zero, this setting has no effect.
- **maxMemoryUsageInMb** – the maximum amount of memory to allow the process to use. If this limit is exceeded the crawler will stop prematurely. If zero, the setting has no effect.
- **maxMemoryUsageCacheTimeInSeconds** – max amount of time before refreshing the value used to determine the amount of memory being used by the process that hosts the crawler instance. This value has no effect if maxMemoryUsageInMb is zero.
- **isForcedLinkParsingEnabled** – sets a value that indicates whether the crawler should parse the page links even if some settings determine that those links will not be crawled.

5.4.3 Disable the indexing of a part of a web page

Web page often contains a lot of information that is duplicated on all pages of the site and useless when searching. Examples of such areas may serve as headers, footers, menus and a variety of navigation elements. To exclude these areas from further processing, the site owner can add the following tags as comments in the html-page:

```
Ordinal text for processing
<!--allowindexing:off-->
This text is not searchable
<!--allowindexing:on-->
Ordinal text again
```

It also supports similar tags used by GSA:

- <!--googleoff: all-->
- <!--googleoff: index-->
- <!--googleon: all-->
- <!--googleon: index-->

This refers to all HTML documents that are stored in the index. However, in case of use of the Web Connector normally your own website will be crawled. Therefore, the owner of the HES software may add to this end the appropriate tags on his website.

5.4.4 Canonical links

Often, the same page of the site is available in several URL. They may differ, for example, at the parameter that shows from where the transition has been made. This makes sense for analysis of the site navigation, but it does not matter when searching for content. At the same time, because of different parameters in links, the crawler perceives them as many pages with the same content.

In such cases, it makes sense to use the mechanism of canonical links that prevents duplicate content issues during indexing of pages by search engines.

To do this enable the option “Use canonical links” in the settings of the Web Connectors and on the site use the page title tag:

```
<link rel="canonical" href="http://example.com/">
```

If some pages will have the same canonical link, the index gets only one (meaning that they are identical in content, or differ insignificantly). The transition from the HES search results will be carried out just on the link in the href attribute.

5.4.5 X-HES-Users headers

The HTTP Protocol allows you to set restrictions on access to certain web resources using the authorization mechanisms. However, HTTP does not give information about the list of users who can access certain documents. By default, all HES users have the same access rights for all documents of Web Connector storage. “X-HES-Users” feature should be used to assign personal access rules for any page and static documents on the site.

There are two Headers: “X-HES-Users-Allowed” and “X-HES-Users-Denied”. These headers must be added to the HTTP response site’s backend.

The names of the users/groups of users for X-HES-Users header can be written in the SID-format or symbolic formats:

- SID (example, S-1-5-21-3255245507-3551417498-1381599987-1124),
- domain\user_name,
- user_name.

Multiple entries within the same X-HES-Users header split ";". One X-HES-Users header can contain names written in different formats.

Response Headers:

```
Accept-Ranges:bytes
Content-Encoding:gzip
Content-Length:132
Content-Type:text/plain
Date:Fri, 02 Sep 2016 07:18:34 GMT
ETag:"0f4993e99e9d11:0"
Last-Modified:Fri, 29 Jul 2016 13:00:56 GMT
Server:Microsoft-IIS/8.5
Vary:Accept-Encoding
X-HES-Users-Allowed:S-1-5-21-3255245507-3551417498-1381599987-
1124;HES\Domänen-Benutzer;HES\j.doe;j.doe
X-HES-Users-Denied:S-1-5-21-3255245507-3551417498-1381599987-1131
X-Powered-By:ASP.NET
```

To properly extraction the data from the X-HES-Users headers, you must consider the following:

Select the mode “Access list mode” that allows you to work with X-HES-Users headers. If the X-headers contain the names in a symbolic format, fill in the “Active Directory credentials” fields (storages settings for the Web Connector) to connect to Active Directory. Note that the field “Username” should be preceded by the domain name.

At the correct filling of fields, symbolic names are converted into SID-format. If the names are not recognized by the system, they will not be saved, and the warning will be displayed on the “DIAGNOSTICS” tab. Reasons for X-HES-Users headers not being recognized, may be as follows:

- Misspelled names of users/group of users in symbolic format (SID format). The names of user/group of users, which were written in SID format, are not verified, so there will not be any message in “DIAGNOSTICS”.
- The header contains the username of another domain or sub-domain.
- Incorrect filling of the settings “Active Directory credentials”. Please note that if you have made changes to this setting, they will take effect only after re-crawling of storage.

6 Useful links

Active directory:

- <https://github.com/hulbee-ag/hes> – HES developers zone (Manuals, HES Connector SDK)
- <https://technet.microsoft.com/en-us/library/dn283324.aspx> – Active Directory Services Overview.
- <https://technet.microsoft.com/en-us/library/hh472160.aspx> – Deploy Active Directory Domain Services (AD DS) in Your Enterprise.
- <https://technet.microsoft.com/en-us/library/jj574166.aspx> – Install a New Windows Server 2012 Active Directory Forest (Level 200).

Organization of backup process:

- <https://technet.microsoft.com/en-US/library/dn390929.aspx> – Windows Server Backup and Storage Pools
- https://en.wikipedia.org/wiki/List_of_backup_software – the list of software for backup (independent vendors, open source).

Elasticsearch

- <https://www.elastic.co/downloads/past-releases> – download page for Elasticsearch v.2.4.*
- <https://www.elastic.co/guide/index.html> – documentations.

7 Known issues

- Uninstalling HES. In case the unsuccessful uninstallation of HES (HES services are still running in the Services applet, a shortcut still present in the Programs and Features applet), try a manual uninstallation scenario:
 1. Stop services Hes.Services.IndexCleaner, Hes.Connectors.Exchange, Hes.Connectors.FileSystem, Hes.Connectors.Web and Hes.Services.ConnectorManager, using “Services” applet.
 2. Run command console (cmd.exe) as Administrator.
 3. Delete services using the following commands:
 - sc delete Hes.Services.IndexCleaner
 - sc delete Hes.Connectors.Exchange
 - sc delete Hes.Connectors.FileSystem
 - sc delete Hes.Connectors.Web
 - sc delete Hes.Services.ConnectorManager
 4. If the services are still visible in the “Services”, reboot the server.
 5. Remove application “hes” of the Default Web Site (IIS) and application pool “hes”, if necessary.
 6. Delete the folder with installed HES. It is C:\Program Files\Hulbee AG\Hulbee Enterprise Search\ folder by default.
 7. Open “Programs and Features” and delete “Hulbee Enterprise Search” shortcut (press “uninstall” and the delete shortcut option will be proposed after this).
- If you need to remove the HES index, it is possible to use Elasticsearch or a tool supplied by Utilities\IndexUtil (help is available if you start the application with the key -help). **Usually, the**

index does not need to be removed. When updating the installer asks whether the index should be rebuilt or the existing one can be used.