



HULBEE ENTERPRISE SEARCH

Installation and setup manual

Modified: 17-June-2016

Version: 2.1.26

Table of Contents

1	ARCHITECTURE	3
1.1	HULBEE ENTERPRISE SEARCH.....	3
1.2	SWISSCOWS COMPANY SEARCH.....	4
2	INITIAL INSTALLATION	5
2.1	FULL INSTALLATION	5
2.1.1	<i>Indexstore</i>	5
2.1.2	<i>Processing server</i>	5
2.1.2.1	Hardware requirements for HES Processing Server	5
2.1.2.2	Software requirements for HES Processing Server	5
2.1.2.3	IFilters setup.....	6
2.1.2.4	Authentication settings	6
2.1.2.5	HES unpacking	6
2.1.2.6	Initial setup.....	7
2.1.2.7	License setup.....	11
2.2	SWISSCOWS COMPANY SEARCH.....	11
2.2.1	<i>Domain setting up</i>	11
2.2.2	<i>File storage</i>	12
3	FIRST RUN	12
4	ADMIN PAGE SETUP	14
4.1	RESTORE OPTIONS.....	14
4.2	THE FIRST SETUP STEPS.....	14
4.3	ADDING CONNECTORS.....	14
4.4	ADDING A STORAGE OF DOCUMENTS.....	14
4.5	ADMINISTRATION AREA	15
4.5.1	<i>Dashboard</i>	16
4.5.2	<i>Connectors</i>	17
4.5.3	<i>Users</i>	17
4.5.4	<i>Logs</i>	18
4.5.5	<i>Unprocessed</i>	18
4.6	SETTINGS.....	19
4.6.1	<i>Authorization</i>	19
4.6.2	<i>Processing</i>	20
4.6.3	<i>Cleaning</i>	23
4.6.4	<i>Logging</i>	24
4.6.5	<i>Import / Export</i>	25
4.7	DATA CLOUD.....	25
4.7.1	<i>Custom queries</i>	25
4.7.2	<i>Stopwords</i>	26
5	CONNECTORS	26
5.1	COMMON TUNES	26
5.1.1	<i>Settings of the connectors via the configuration files</i>	26
5.1.2	<i>Settings of the connectors in the admin area</i>	28
5.2	FILESYSTEM CONNECTOR.....	29
5.2.1	<i>File connector settings</i>	30

5.3	WEB-CONNECTOR.....	30
5.3.1	Settings of the web connector	30
5.3.2	Fine tuning of the web connector	32
5.3.3	Disable the indexing of a part of a web page	32
6	USEFUL LINKS	33
7	KNOWN ISSUES	33

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. For the use of HES 2.1 two types of resources are available “from the box”: the enterprise file system and web resources. 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 each source are added to the system and referred 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		✓	

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 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.

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	✓	✓	

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 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 Swisscows Company Search

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.

Swisscows Company Search contains versions notable both for Hardware capabilities and for limitation of number of users.

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).

2 Initial Installation

2.1 Full installation

2.1.1 Indexstore

Elasticsearch v.1.7 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 operating system different from Windows, but supporting JRE (GNU/Linux, Solaris, etc).

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

After installation in accordance with Elastic website recommendations, it is necessary to make post-installation settings in Elasticsearch configuration (elasticsearch.yml file). Add the following lines at the end of the file:

```
script.inline: on
script.indexed: on
```

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.

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-Security	IIS-NetFxExtensibility45	IIS-HttpCompressionStatic	IIS-WebSockets
		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 user name and password of the current Windows user, requires the inclusion of the appropriate IIS configuration: In the IIS Manager window select a site with HES (usually it is the Default Web Site). In the features panel select “Authentication”. Here it is necessary to include the following items: “Anonymous Authentication” and “Windows Authentication”. For “Windows Authentication” in the section “Action”> “Enable”> “Providers ...” select the appropriate “Negotiate” provider.

2.1.2.5 *HES unpacking*

HES applications pack has a name like **HES.2.1.XX.XXXXX.zip (XX being the numbers of your specific version)**. It contains the following components:

- Connectors
 - 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).

2.1.2.6 Initial setup

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

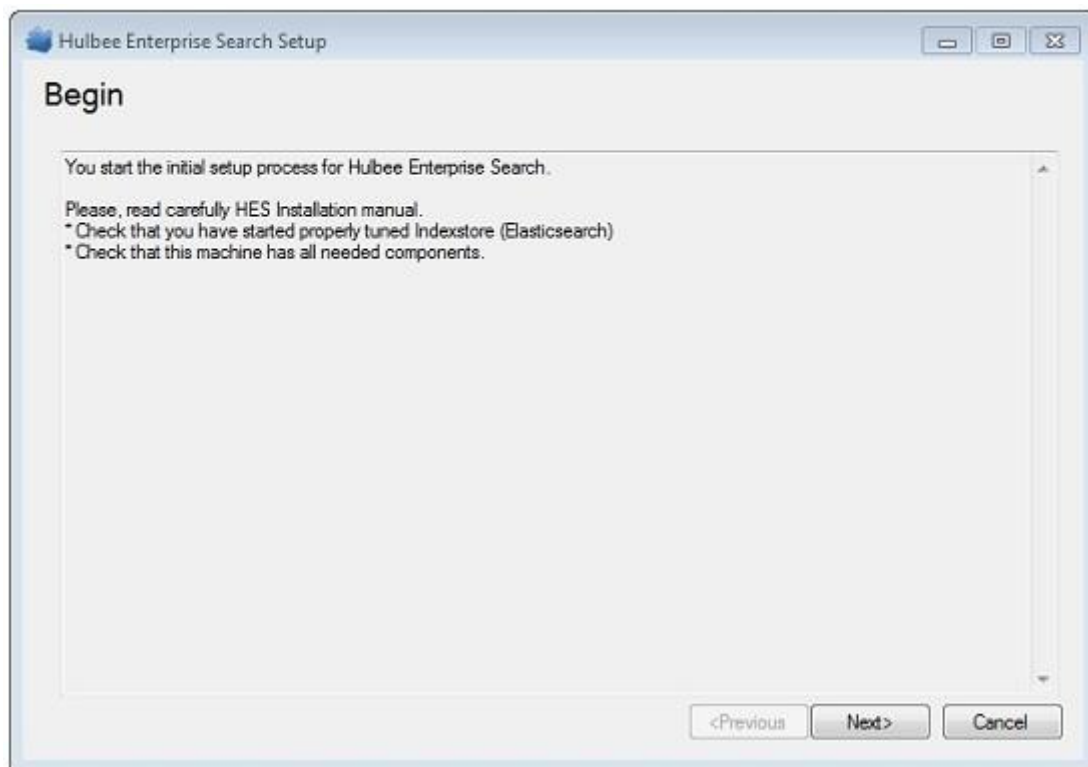


Fig. 1. HES Setup – start of Setup.

Click the "Next" button to take step 1.

1. Step 1.

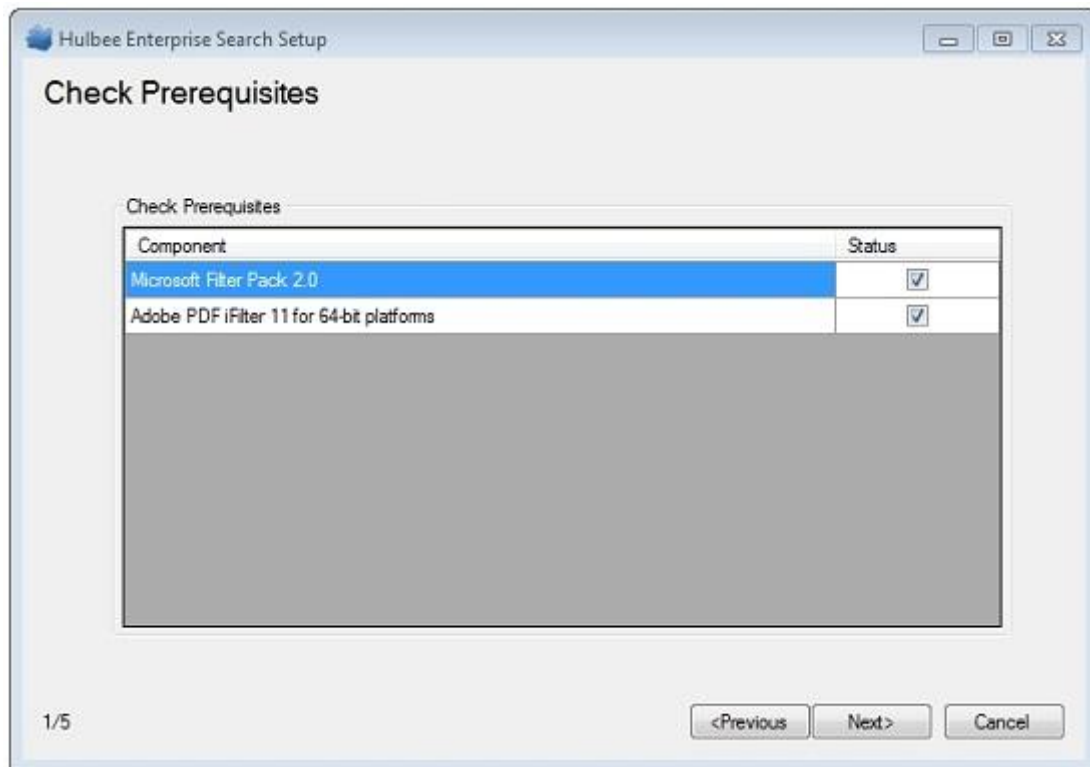


Fig. 2. Step 1 of HES Setup.

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

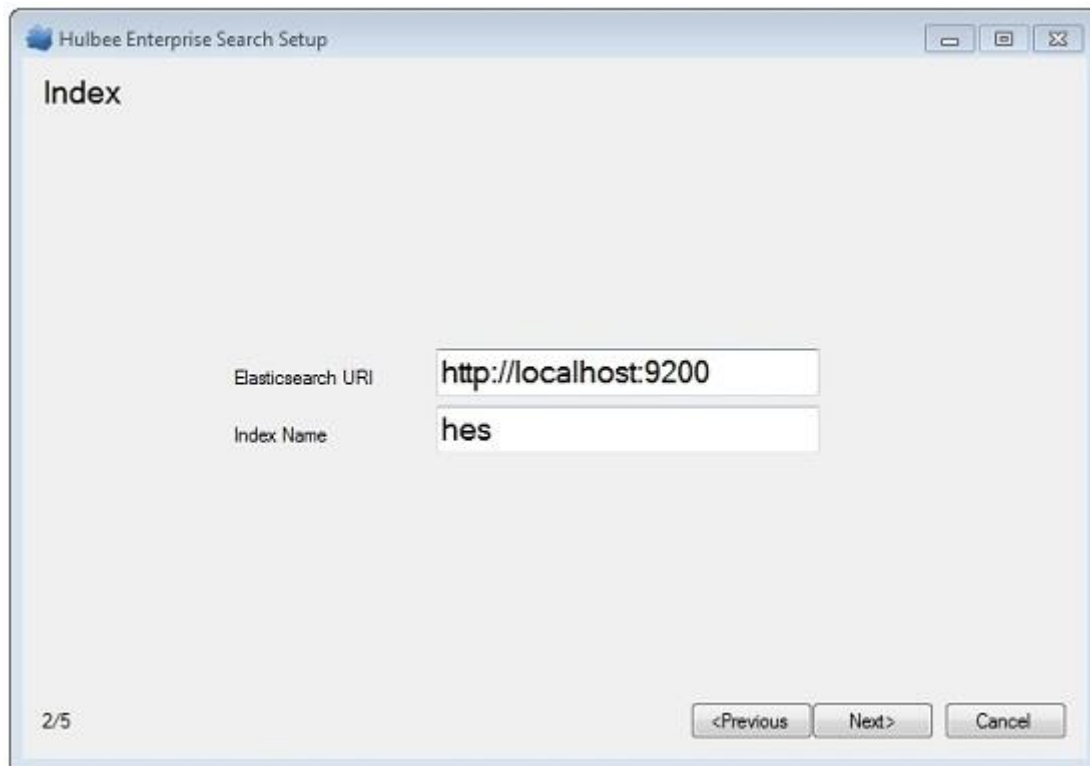


Fig. 3. Step 2 of HES Setup.

3. Step 3.

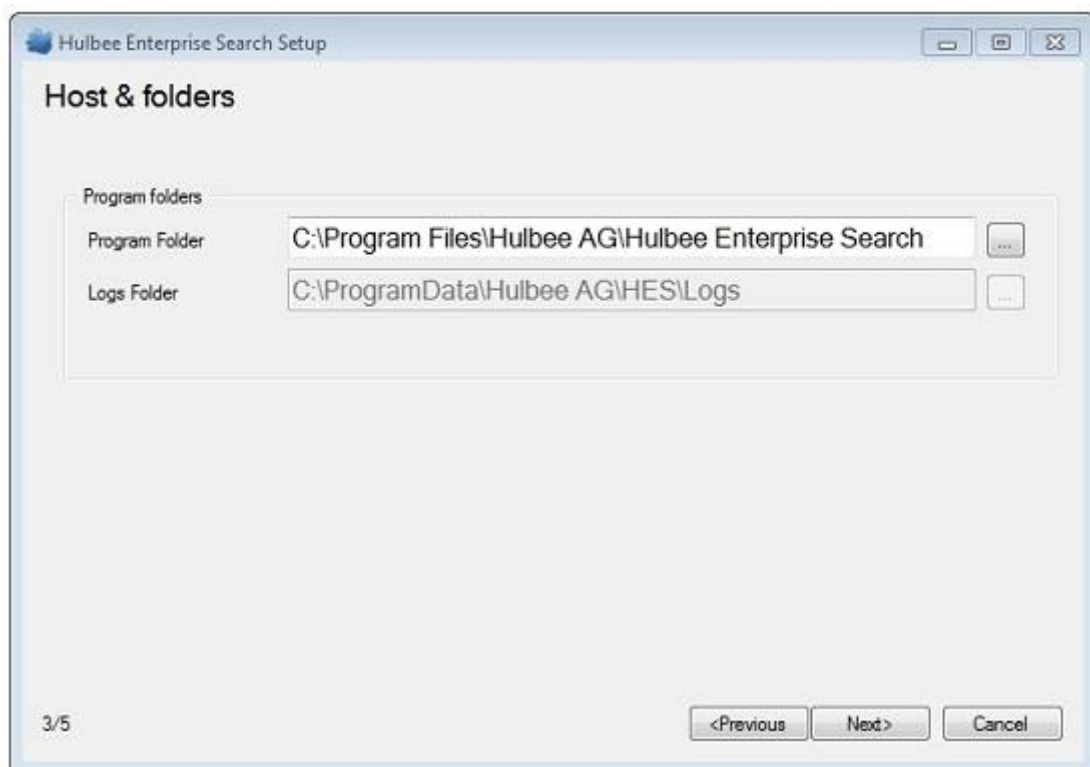
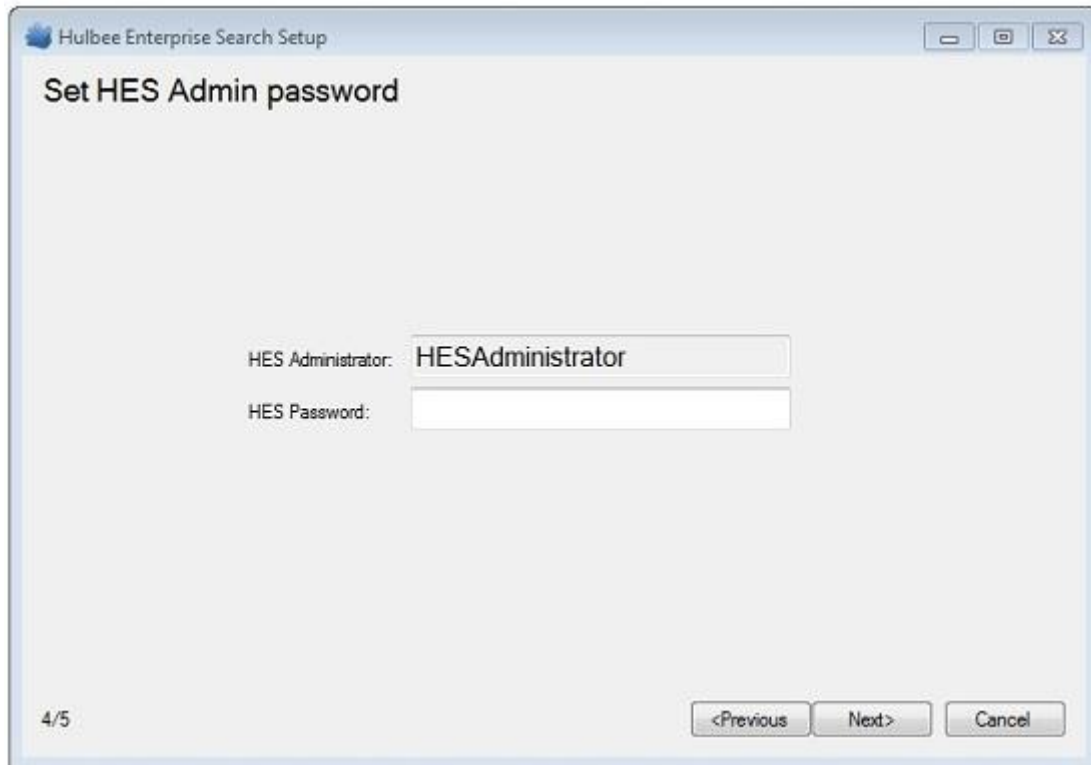


Fig. 4. Step 3 of HES Setup.

“Program folder” and **“Logs folder”** can be left with default values.

4. Step 4.

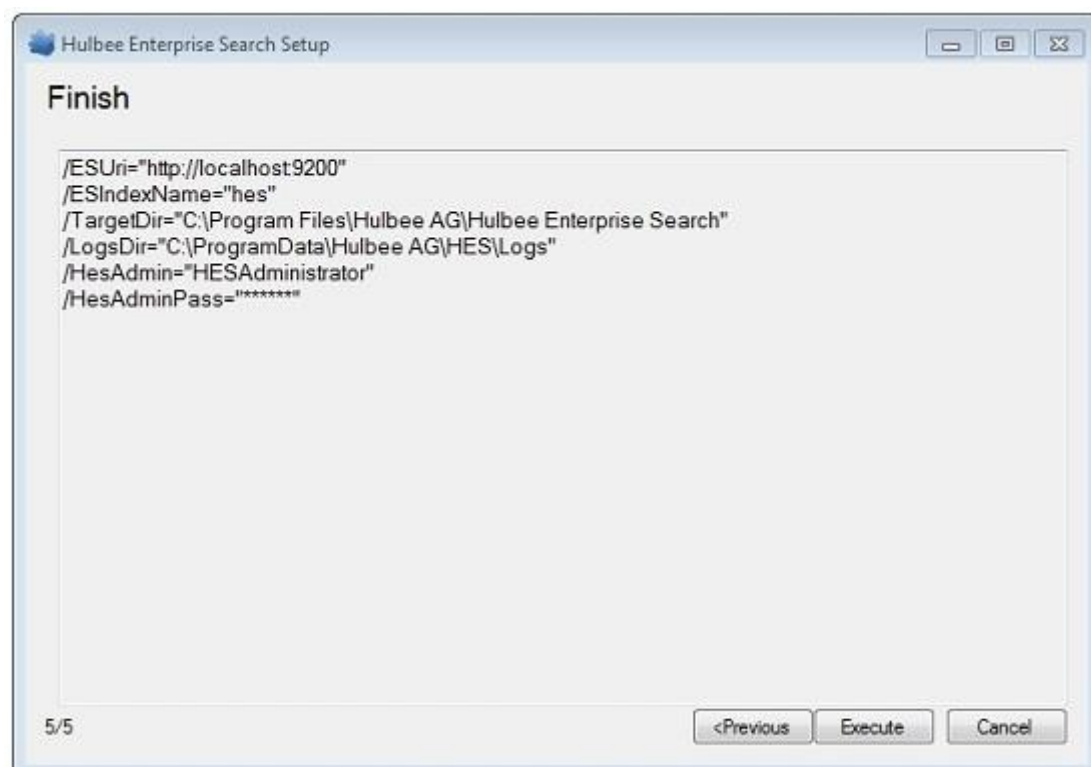
In this step, the HES administrator password is entered. This user is very special – he is not a user of the Active Directory or the current machine and he is needed in order to be able to authenticate on the HES software that is not set up yet. Only he has access to the admin panel of HES and can make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



The screenshot shows the 'Set HES Admin password' window of the 'Hulbee Enterprise Search Setup' wizard. The window has a title bar with the text 'Hulbee Enterprise Search Setup' and standard Windows window controls. The main title is 'Set HES Admin password'. Below the title, there are two input fields: 'HES Administrator:' with the text 'HESAdministrator' entered, and 'HES Password:' which is empty. At the bottom left, it says '4/5'. At the bottom right, there are three buttons: '<Previous', 'Next>', and 'Cancel'.

Fig. 5. Step 4 of HES Setup.

5. Step 5.



The screenshot shows the 'Finish' window of the 'Hulbee Enterprise Search Setup' wizard. The window has a title bar with the text 'Hulbee Enterprise Search Setup' and standard Windows window controls. The main title is 'Finish'. Below the title, there is a text area containing the following configuration details:
/ESUri="http://localhost:9200"
/ESIndexName="hes"
/TargetDir="C:\Program Files\Hulbee AG\Hulbee Enterprise Search"
/LogsDir="C:\ProgramData\Hulbee AG\HES\Logs"
/HesAdmin="HESAdministrator"
/HesAdminPass="*****"
At the bottom left, it says '5/5'. At the bottom right, there are three buttons: '<Previous', 'Execute', and 'Cancel'.

Fig. 6. Step 5 of HES Setup.

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

6. Final step.

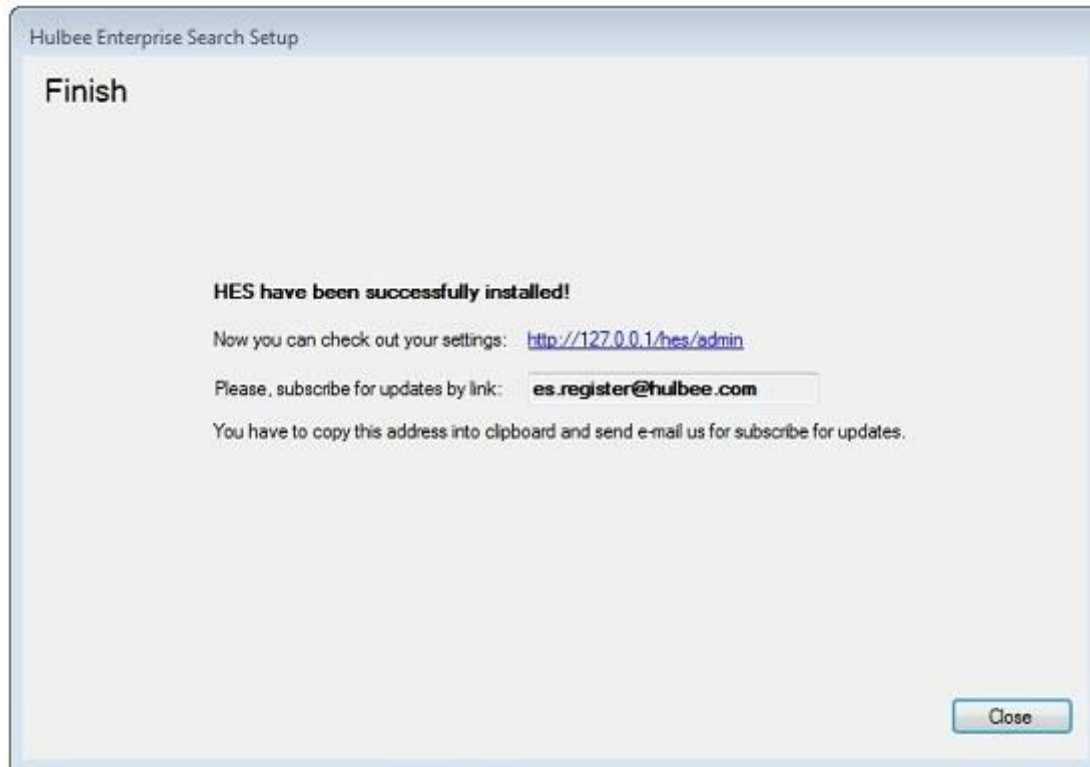


Fig. 7. HES Setup – setup finish.

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).

2.1.2.7 License setup

In addition to the product, a personal license file is provided. If the file was not made available, you can 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.

2.2 Swisscows Company Search

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

- Joining the Swisscows Company Search 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 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's password use standard procedure.

To change the HES administrator's 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 server with sufficient storage capacity disk array, or to set additional disks in Appliance Server.

Notice! The file storage should be available at 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 logging in, fill out an authorization form:

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:

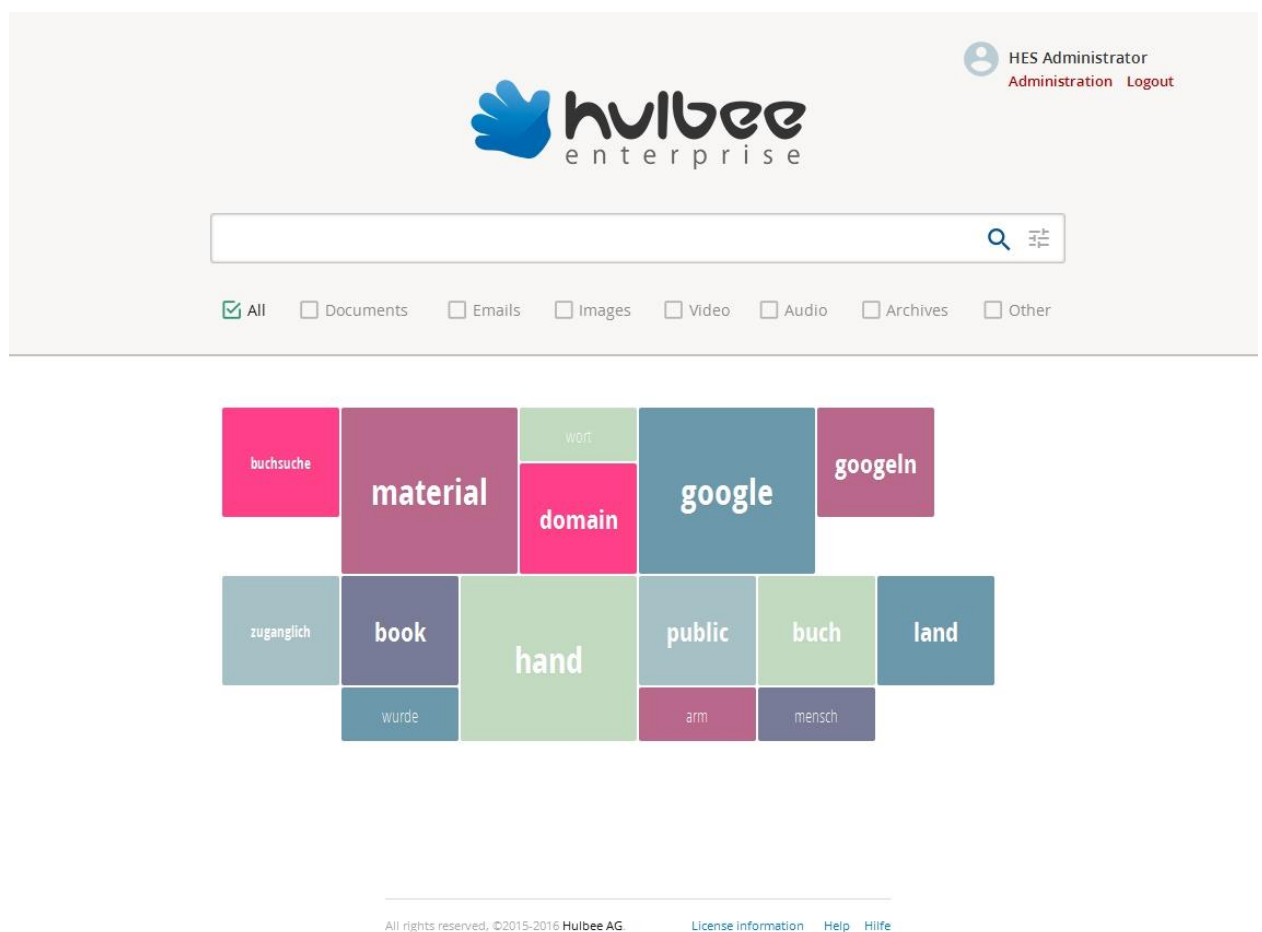


Fig. 9. HES start page.

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".

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. Add connectors (see 4.3).
3. Add storages of documents (see 4.4).

4.3 Adding connectors

When you first sign in no connectors are available in the list. HES package includes two types of connectors: for the connection to the file system of the company and to the web resources. There is the possibility to add your own custom connectors using the HES Connector API (see 6 – Useful links). To register the connectors, perform the following 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 and 5.3). The port on which the custom connector are registered must be requested from the developer.
2. Click the "Register" button.

Notice! Check that the connector service was launched during the execution of the connection.

4.4 Adding a storage of documents

Once the necessary connectors have been added to the HES software, the storages, where the documents are kept, need to be added. Storage is a place where documents are stored in the same way. This may be: network file sharing, website, CRM system, ERP, mail servers, etc. 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). It is not necessary to register several connectors of the same type in the case of standard connectors from the package.

To add storage, select “Details” button in “Connector”.

Connectors

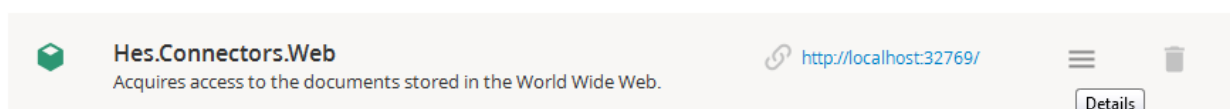


Fig. 10 Connector area, details.

Next, select “Settings” tab, where operations such as adding, deleting, configuring of storages can be performed. Details of these procedures are described in sections 5.2 and 5.3.

4.5 Administration Area

Pages of admin panel contain navigation area and workspace. Navigation area contains links. Following these links, you will be shifted to different subsections of admin panel. It is the same for all pages. The workspace contains elements for options edit and display of system operation information.

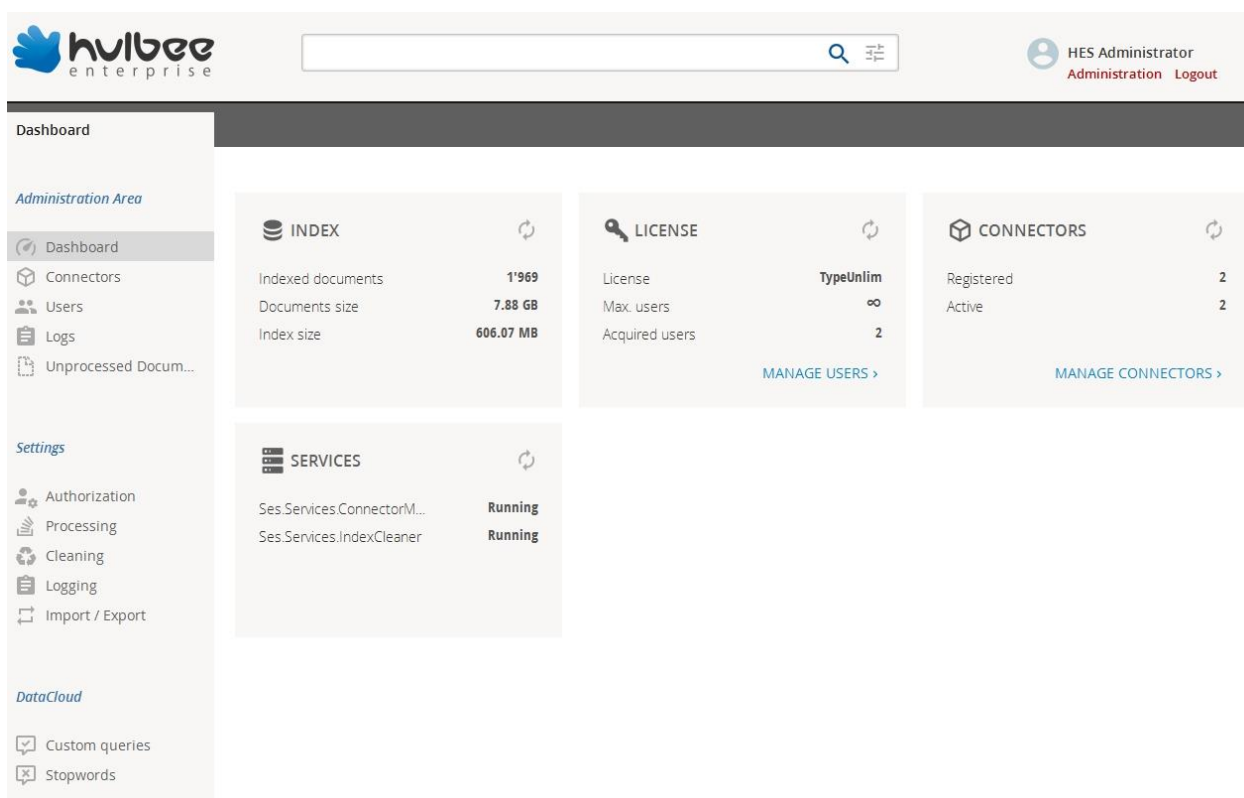


Fig. 11. Administration area. General view.

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

4.5.1 Dashboard

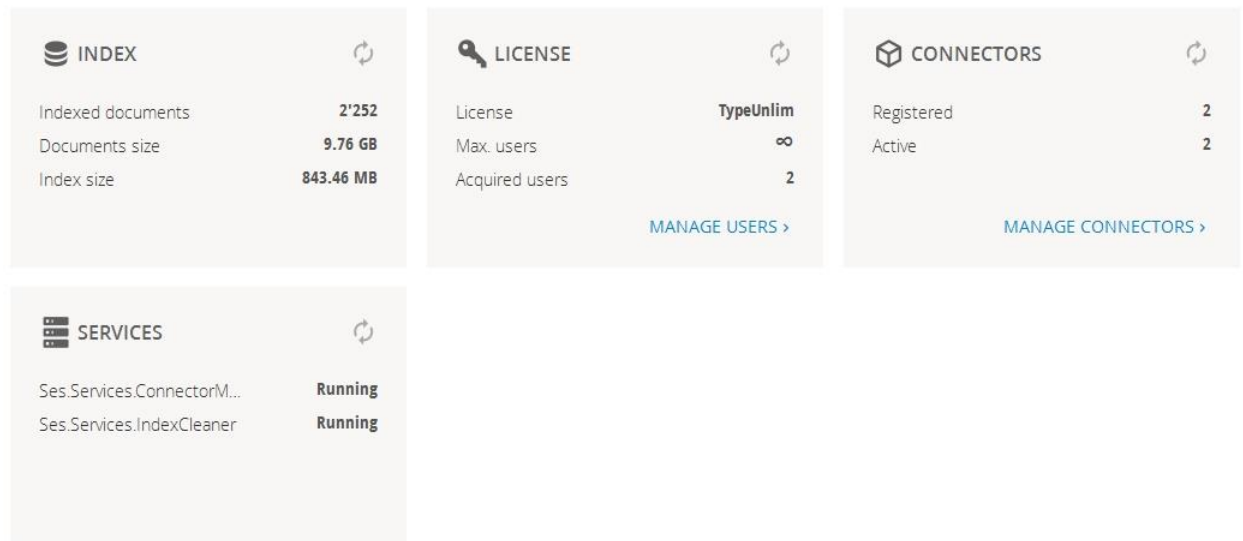


Fig. 12. Dashboard page.

You can see different kinds of statistics on this page:

- **INDEXED DOCUMENTS.** The number of indexed documents. It can differ from the total number of files in storage, 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 index. The sum of their binary sizes is meant. It can differ from the space occupied in the file storage by the same reasons as for preceding item.
- **INDEX SIZE.** The size of Elasticsearch index on the disk.
- **LICENSE {id}.** This 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

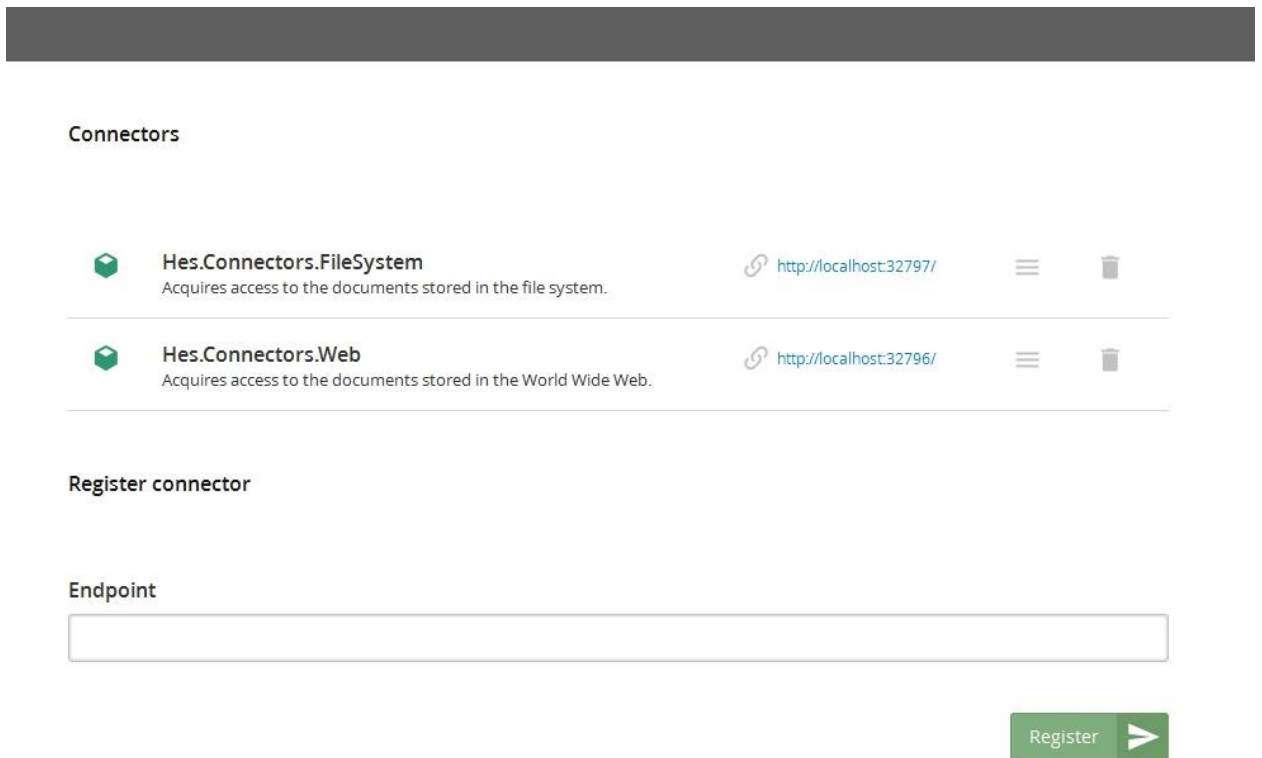


Fig. 13 Connectors page.

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.

Existing connectors can be removed 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 52			
Critical, Error, Warning, Information, ...			
Ses.Services.ConnectorManager, Ses....			
TIMESTAMP	EVENT	MESSAGE	PROCESS
Today, 11:12:27	Start	Started processing of the feed `Users` (Hes.Connectors.FileSystem).	Ses.Services.ConnectorManager
Today, 11:12:26	Start	Started processing of the feed `Common` (Hes.Connectors.FileSystem).	Ses.Services.ConnectorManager
Today, 11:12:23	Stop	Finished processing of the feed `Users` (Hes.Connectors.FileSystem).	Ses.Services.ConnectorManager
Today, 11:12:22	Warning	Couldn't find any active feed on connector Hes.Connectors.Web.	Ses.Services.ConnectorManager
Today, 11:12:22	Warning	Couldn't find any active feed on connector Hes.Connectors.Web.	Ses.Services.ConnectorManager
Today, 11:12:22	Warning	Couldn't find any active feed on connector Hes.Connectors.Web.	Ses.Services.ConnectorManager
Today, 11:12:22	Warning	Couldn't find any active feed on connector Hes.Connectors.Web.	Ses.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 list Events.

4.5.5 Unprocessed

Page 1 of 1					
Unprocessed Documents					
<i>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.</i>					
	URI	STATUS	ERROR		
	file:///storage/files/QA/TEST-DATA/Archives/7z/audio/wav.7Z	Aborted	File is to large to be processed.		
	file:///storage/files/QA/TEST-DATA/FileSize%20-%20text/126MB.pdf	Aborted	File is to large to be processed.		
	file:///storage/files/QA/TEST-DATA/FileSize%20-%20text/150MB.xml	Aborted	File is to large to be processed.		

Fig. 16 Unprocessed page.

The list of documents that have not been processed due to some errors. For a single unprocessed file the time is defined during which an attempt was made to process it and the reason is given why no file processing was carried out. You can trigger some items to try process them once again or schedule processing with extended limits for the extremely important documents.

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.

hes.hulbee.local

NetBIOS Domain Name

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

HES

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_user

Password

The password used to connect to the store.

••••••••

Access Rules

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

<input type="checkbox"/>	IDENTITY	ACCESS	
+	User or group SAMAccountName, e.g. DOMAIN\Everyone	Allow	▾

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, a 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.

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 with the user from Active Directory, whose data will be specified in the fields “Username” and “Password”. This field is optional. The default value for this field is an empty string. This 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 granted or denied can be accurately determined. If the field remains empty, the access applies for all users of the domain.

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



The screenshot shows a configuration page for the 'Processing' section. It contains three settings:

- Concurrent workers count**: A description states 'The maximum number of concurrent operations that can be run by the processing engine.' The input field contains the value '2'.
- Watchdog timeout**: A description states '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.' The input field contains the value '00:05:15'.
- Idle timeout**: A description states '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.' The input field contains the value '00:05:15'.

Fig. 18 Processing page.

Select “Processing” area and tune following fields:

- **“Concurrent workers count”**: 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 shouldn’t be greater than the number of cores in the Processing Server. For configuration of Swisscows Company Search 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 can be viewed in a preview 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.

- **“Maximum document size”**: The documents whose size in bytes exceeds the given value will not be processed.
- **“Maximum 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.
- **“Maximum attachments size”**: Maximum size of attached files. Having reached this size, archive or mail message processing stops.

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 extensions”**: The files of some types are large and fall within size limitation. But their processing is quite easy, as 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



The screenshot shows a web interface for configuring cleaning settings. It features two main sections: 'Batch size' and 'Iteration timeout'. The 'Batch size' section has a text input field containing '1000' and a small blue button with up and down arrows. The 'Iteration timeout' section has a text input field containing '00:00:10'. Below these fields is a green 'Save Settings' button with a white checkmark icon.

Batch size
The number of documents to be processed per one iteration.

1000

Iteration timeout
The timeout between iterations. The value must be in format "HH:MM:SS", e.g. 00:05:00 - five minutes.

00:00:10

Save Settings ✓

Fig. 21 Cleaning page.

Select "Cleaning" area and tune following fields:

- **"Batch size"**: The number of documents to be processed per one iteration. Recommended value – 1000. The estimated value of the number of documents that are scanned in one step. The real value of each portion may vary slightly.
- **"Iteration timeout"**: The timeout between iterations. In most cases, 10 seconds is a proper time. When we speak about processing of millions and tens of millions documents, we can decrease this time (it increases load on Processing server and Indexstore) to delete already deleted documents from index (or in the case of documents storage path change).

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

4.6.4 Logging

The screenshot shows a web interface for logging settings. It includes a 'Trace level' dropdown set to 'Information', an 'Automatically remove logs older than' dropdown set to 'Week', and a 'Save Settings' button with a checkmark. Below these are 'Event types' (set to 'Critical, Error, Warning, Information, Verbose, Start, Stop') and a 'Before date' field set to '25.05.2016' with a calendar icon. A 'Delete' button with a trash icon is at the bottom right.

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

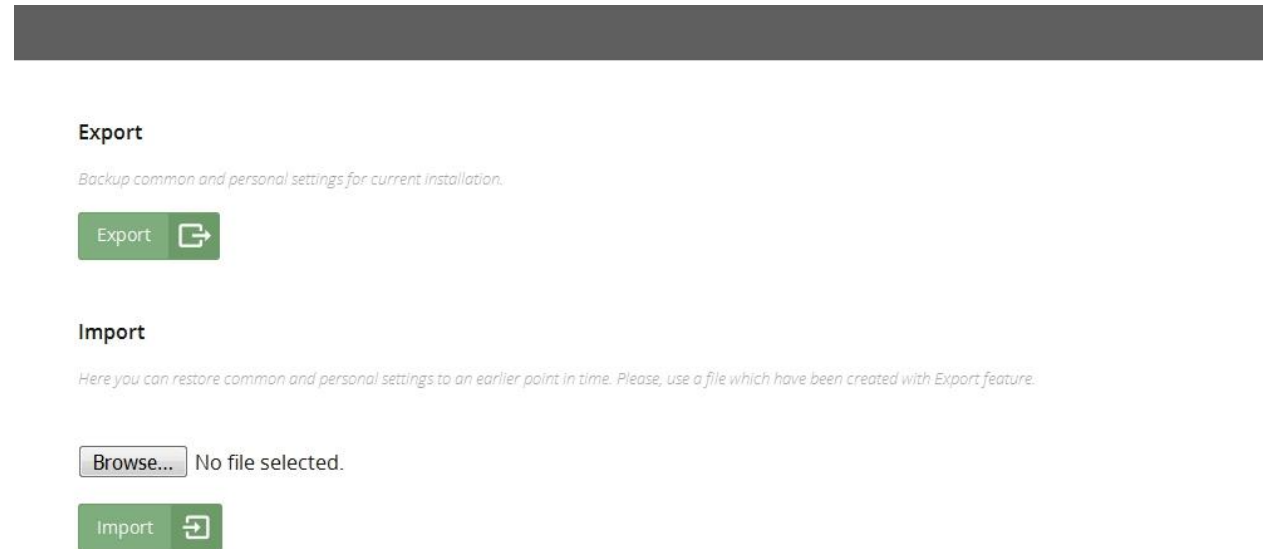


Fig. 23 Import / Export page.

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.

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

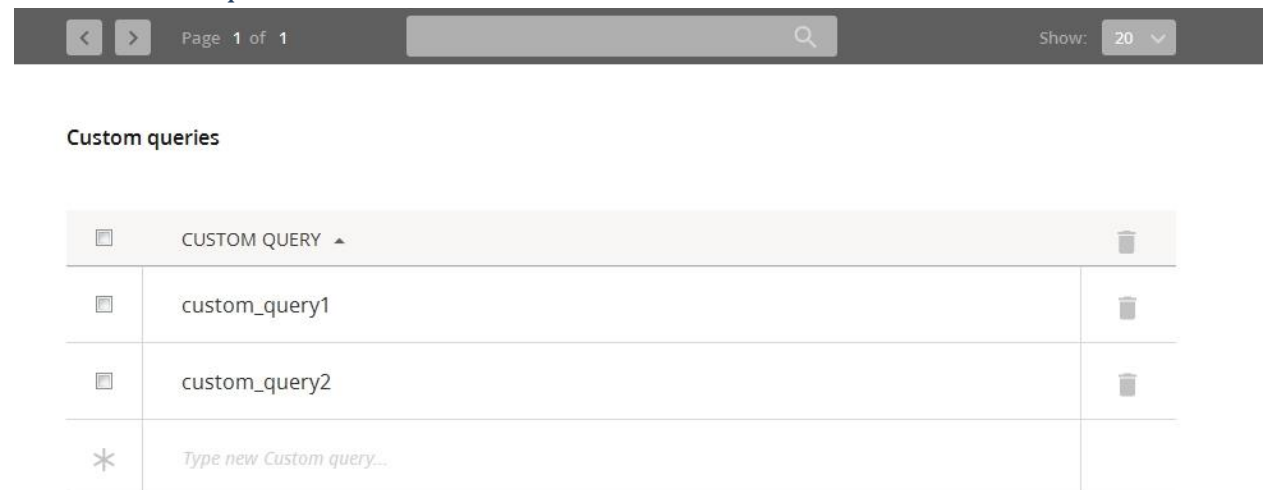


Fig. 24 Custom queries page.

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

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

4.7.2 Stopwords

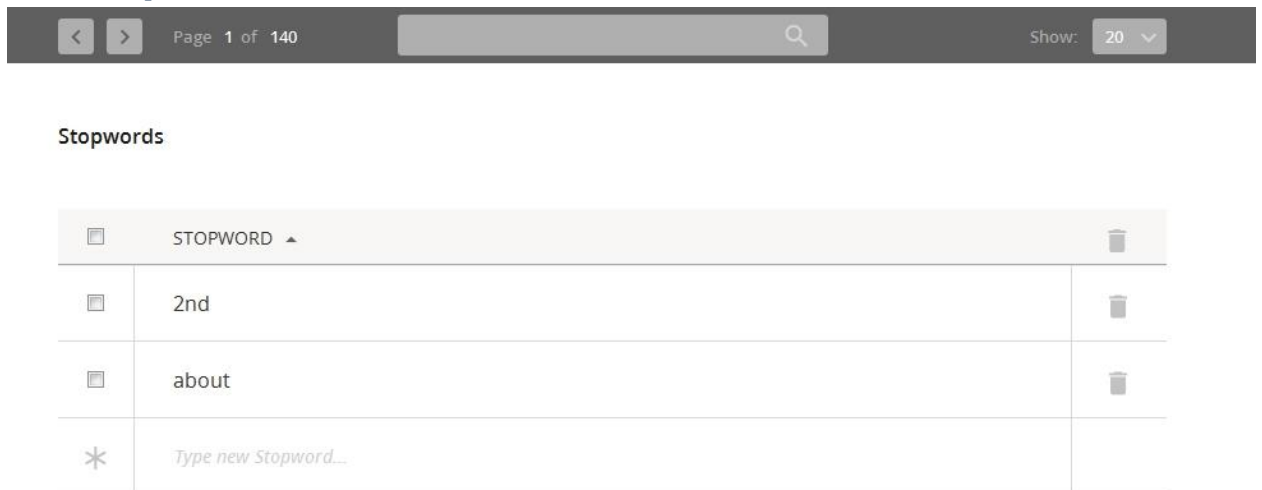


Fig. 25 Stopwords page.

Using “Stopwords” tab you can indicate words, which you do not want to see in 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.

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 perform the following steps:

- 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

- 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 step 1.

- 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 step 1, ** – any valid GUID (for its generation can be used, for example, <https://www.guidgenerator.com/>).

Authentication can be configured using “authentication” area:

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

  <basic enabled="true or false" username="allowed user name"
    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, which allows you to go to the tabs (the settings for connectors).

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

← Back to connectors list

Hes.Connectors.FileSystem
<http://localhost:32797/>

GENERAL

SETTINGS

FEEDS

STATISTICS

DIAGNOSTICS

☒ **Enable data processing**
If checked the documents provided by the connector will be added/updated to index; otherwise it will be treated by IndexCleaner only.

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

Authentication method
The authentication method used when performing requests to connector.

None

Save Settings

✓

Fig. 26 Connectors. “GENERAL” tab.

- **“Enable data processing”**: In switched state means that the documents in the index will be indexed, updated, added. If the mark is removed, documents are only checked for removal from the index by the IndexCleaner, but an active traversal of the storage will not be made.
- **“Display Name”**: Name of the connector, which is set by the administrator. It is displayed on the user's page in the filter and on the page of the advanced search.
- **“Authentication method”**: 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 has access to the connector.

In the “SETTINGS” tab settings are displayed (section 5.2.1 and 5.3.1), which are different for the respective types of connectors and are described in the documentation.

The list of settings on the “FEEDS” tab is available as soon as they are opened in one of the storage lists (Fig. 27).

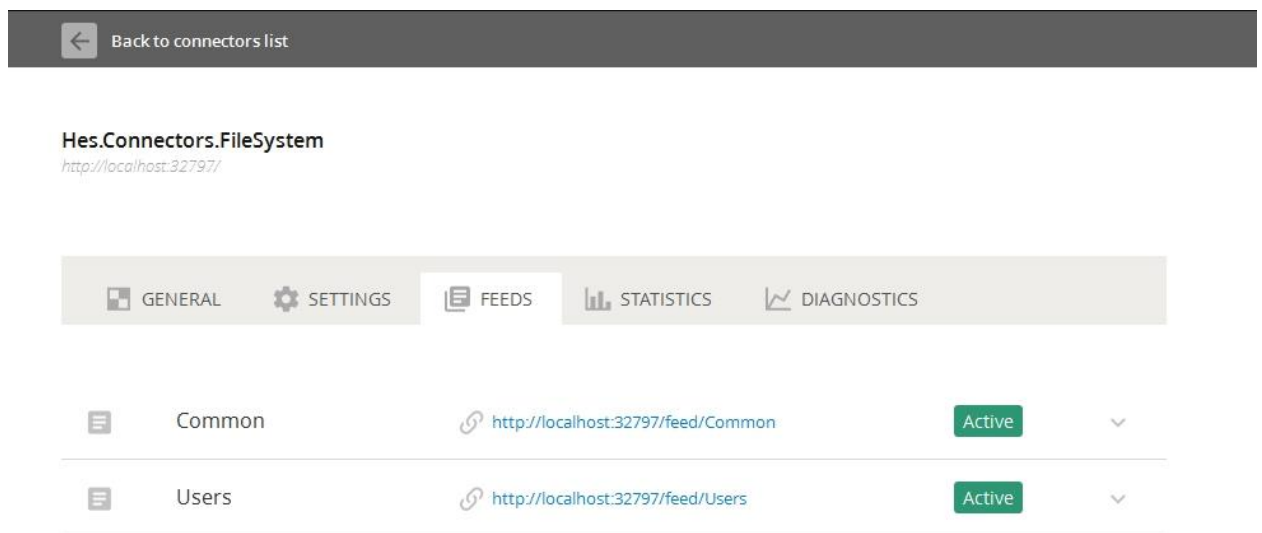


Fig. 27 Connectors. “FEEDS” tab.

The following settings from the traversal of the storages are displayed:

- **“Enable data processing”**: If checked the documents provided by the feed will be added/updated to index; otherwise it will be treated by IndexCleaner only.
- **“Display Name”**: The name, which will be displayed for the feed in filters, metadata, etc.
- **“Polling batch size”**: The count of documents to fetch per traversing request.
- **“Polling interval”**: Minimal timeout between traversing requests.
- **“Web requests interval”**: Minimal timeout between web request to obtain http(s)/ftp contents.

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

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

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.

When you add new storages, you need to specify their settings. The settings for the file connector are available as soon as they are opened in one of the storage lists. Settings include:

- **"Storage name"**.
- **"Scan location"**: Path on the local machine or in the local area network, which should be scanned.
- **"Scan/watch files wildcard"**: The search pattern to match against the names of files in scan location. You can enter only extension or only the name (part of the name) to search the file. Example: *.tmp or doc*.* (doc*) or document.*. Entering multiple wildcard at the same time through a comma without gaps.
- **"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 patterns is written into a separate line, with a comma.
- **"Access credentials"**: You can specify the identity credentials, which will be used to access the storage.
- **"Domain", "Username", "Password"**: In these fields data of the user is entered, who has access to the repository.

5.3 Web-connector

This crawler traverses the pages in the local and global network. The web connector has extensive settings of the crawlers. All documents that are found under this crawler have the same settings of visibility for users within the memory (setup Access list allowed and Access list denied).

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

5.3.1 Settings of the web connector

The settings for the connector are available as soon as they are opened in one of the options in the list of storages. The settings for the Web connector include:

- **"Storage name"**.
- **"Url"**: Url of internet resource's start page 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. This value is required.
- **"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 on "Enable external page crawling" and "Enabled

external links crawling", processing of pages and documents is made not only in the domain specified on the home page. If zero, this settings has no effect.

- **"Max 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.
- **"Enable external page crawling":** Whether pages external to the root uri should be crawled.
- **"Enabled external page links crawling":** Whether pages external to the root uri should have their links crawled. This setting is useful only if the previous setting is enabled.
- **"Http request timeout (seconds)":** Time limit for handling the resource.
- **"Http request maximum auto redirects":** The maximum number of redirects that the request follows. If zero, this settings has no effect.
- **"Enabled http request auto redirects":** If a link with redirects exists, gets or sets a value that indicate whether the request should follow the redirection. The link itself is not added.
- **"Enabled sending 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 settings is useful to crawl sites with an invalid or expires SSL certificate. Useful in the processing of https resources.
- **"Max crawl depth":** Maximum levels below root page to crawl. If value is 0, the homepage will be crawled but none of its links will be crawled.
- **"Max retry count":** The max number of retries for processing the file, if the file could not be processed in the given time. If the value is 0, no retries will be made.
- **"Min retry delay (milliseconds)":** The minimum delay between a failed http request and the next attempt to re-access the file.
- **"Enabled respect robots.txt":** Whether the crawler should retrieve and respect the robots.txt file rules.
- **"Enabled respect meta robots no follow":** Whether the crawler should ignore links on pages that have meta-tag of nofollow: <meta name="robots" content="nofollow" />.
- **"Enabled respect HttpX Robots tag header no follow":** Whether the crawler should ignore links on pages that have an http X-Robots-Tag header of nofollow.
- **"Enabled respect anchor rel no follow":** Whether the crawler should ignore links on pages that have rel-attribute of nofollow: .
- **"Enabled ignoring robots.txt if root disallowed":** When this option is enabled, robots.txt is ignored.
- **"Robots.txt user agent":** Allows specification of Robots agent if robots.txt rules are selected. If the option is not selected, the general rules apply.
- **"Max 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.
- **"Min crawl delay per domain (milliseconds)":** The number of milliseconds to wait in between http requests to the same domain. Setting is needed in order to not create too much strain on the site by crawling. If, for example, 500 milliseconds is specified, then it would mean that no more than 2 pages are requested from the site per second.
- **"Additional headers":** Additional http-headers for the site are written in the format key: value. Each individual title is recorded in a new line. It is necessary in rare cases.

- **“Min generations before deleting”**: The number of attempts to download a document for review, which is stored in the database by crawling. Before it can be deleted from the list of documents available from the website.
- **“Pause after cycle (seconds)”**: Delay between cycles of traversing the resource.
- **“Access list enabled”**: If the option is disabled, access to documents from a given repository is given to all users of an instance of HES; otherwise rules for the user are generated, Access list allowed / denied.
- **“Access list allowed”**: A list of users and groups that have access to the documents in memory. A part may be denied access using the settings from the next paragraph.
- **“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 the previous point. That is, the list of the denial takes precedence over the list of permissions.
- **“Taboo rules”**: A list of regular expressions. If the link is subject to at least one rule, it will not be indexed and can further be removed. Every single regular expression is written into a separate line. This way the crawler does not crawl unimportant parts of the site (or those where he can go in cycles, avoiding an infinite number of pages). Another application setting is splitting one site to multiple repositories. 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 the following value actpos=2, must be placed in another storage with other access settings. To do this, we enter in Taboo rules of the primary storage “actpos=2”. It would prohibit getting such pages to the main storage. And for this very special part of the site storage must be added to its website and its own rule “.*actpos=(?!2).*”. It will work if this option is not there.

5.3.2 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:

- maxPageSizeInBytes – if the size of the download page is exceeded, then “abot” will not download it, and searches for the links. But they will still enter to the index.
- downloadableContentTypes – first, the titles of the pages are downloaded and their content-type is checked. If they match, the site is downloaded for the following search of the links on the new page (the work of the crawler).

5.3.3 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 applies 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 crawling. Therefore, the owner of the HES-copy may add to this end the appropriate tags on his website.

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 1.7.*
- <https://www.elastic.co/guide/index.html> – documentations.

7 Known issues

- To install Elasticsearch v.1.7 (see 2.1.1) java 8 must be installed, update version 73. If you are using HES 2.1, in order to avoid malfunctions, do not run the update of java versions.
- 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.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.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.**