



HULBEE ENTERPRISE SEARCH

Installation und Setup-Handbuch

Modified: 17-Dez-2015

Version: 1.7.15

Inhaltsverzeichnis

1	ARCHITEKTUR	3
1.1	HULBEE ENTERPRISE SEARCH	3
1.2	SWISSCOWS COMPANY SEARCH	4
2	ERSTINSTALLATION	4
2.1	VOLLSTÄNDIGE INSTALLATION	4
2.1.1	Indexstore	4
2.1.2	Verarbeitungsserver	5
2.1.2.1	Hardware-Voraussetzungen für HES Processing Server	5
2.1.2.2	Software-Voraussetzungen für HES Processing Server	5
2.1.2.3	Domain und Benutzer-Setup	6
2.1.2.4	IFilters setup	6
2.1.2.5	HES Auspacken	6
2.1.2.6	Ersteinrichtung	6
2.2	SWISSCOWS COMPANY SEARCH	12
2.2.1	Einrichtung einer Domain	12
2.2.2	Dateispeicher	13
2.2.3	Einstellungen	13
3	ÜBERPRÜFUNG DER INSTALLATION	13
4	EINRICHTUNG DER ADMIN SEITE	14
4.1	WIEDERHERSTELLUNGSOPTIONEN	14
4.2	ADMINISTRATIONSBEREICH	15
4.2.1	Dashboard	16
4.2.2	Logs	16
4.2.3	Benutzer	17
4.3	EINSTELLUNGEN	18
4.3.1	Crawling	18
4.3.2	Cleaning	21
4.3.3	Logging	22
4.3.4	Import / Export	23
4.4	DATA CLOUD	23
4.4.1	Benutzerdefinierte Abfragen	23
4.4.2	Stoppwörter	24
5	HILFREICHE LINKS	24
6	BEKANNTE PROBLEME	25

1 Architektur

1.1 Hulbee Enterprise Search

Hulbee Enterprise Search (HES)¹ ist für Nutzer von MS Active Directory vorgesehen. Die aktuelle Version ist für die Suche in Dateispeichern unter Berücksichtigung der Rechte von Nutzern an den einen oder anderen Dateien vorgesehen. Für eine Einbeziehung von Dateien in die Suche muss der Nutzer über Rechte, jedoch mindestens über Leserechte verfügen.

Es werden folgende Dateiformate unterstützt:

Dateityp	Dateierweiterung	Extrahieren von Text	Extrahieren von Meta-Tags	Extrahieren von eingebetteten Dateien
Textdateien	txt, rtf, doc/dot, odt, wri, sxw	✓		
	docx/docm/dotx	✓	✓	
Verarbeitungsdateien	pdf	✓	✓	
	xps	✓		
Hypertextdateien	html, htm, xml	✓		
	mht, shtml	✓	✓	
Tabellen	xsl, xslt, xls, ods, csv	✓		
	xlsx	✓	✓	
Präsentationen	pptx	✓	✓	
	ppt, pps, odp	✓		
Grafische Dateien	bmp, jpg/jpeg, png, jfif, tif, tiff, jpe		✓	
E-mail	msg, eml	✓	✓	✓
Archive	zip, rar, 7zip			✓
Media	avi, mp3, mp4, wav, m4a, wma, wmv,ogg, flac, mkv, ape, mpc		✓	
Source Code und Scripting	cs, vb, js, csproj, h, c, cpp, vbs,vcproj, vbproj, pl, sql, bat, cmd	✓		
	css	✓	✓	

Als Nutzeroberfläche kann jeder beliebige Browser (Mozilla Firefox, Chrome u.a.) mit offenem Link zur Suchseite im Intranet verwendet werden. In ihm werden die Suche selbst und Änderungen des Nutzerprofils vorgenommen. Auch Administratoren können unter Nutzung der Administratorenleiste viele Einstellungen vornehmen.

¹ Einige der Module haben Swisscows oder SES im Dateinamen oder Texten. Es ist der alte Name des Projekts und ist ein Synonym für Hulbee Enterprise Search oder HES.

Zum Öffnen der gefundenen Dateien ist es erforderlich, dass der Nutzer mit seinem echten Account in einem lokalen Netz mit Dateispeicher arbeitet. Das Öffnen der Dateien erfolgt mithilfe der HES Desktop Manager Utility, die auf dem Computer des Kunden installiert sein muss. Der Nutzer muss über die erforderlichen Rechte für die Installation von Software verfügen oder sich hierfür an seinen Administrator wenden.

HES-Software-Komplex besteht aus zwei Hauptteilen:

- Indexstore (Linux-Server mit installiertem und konfiguriertem Elasticsearch).
- Processing server/ Verarbeitungsserver (Windows-Server mit allen anderen Komponenten).

Bei kleinen Filestores, könnten diese Server mit den einzelnen Windows-Server-Maschinen kombiniert werden. Grosse Filestores können eine Anpassung von Elasticsearch Cluster mit einer Anzahl von Servern erfordern.

1.2 Swisscows Company Search

Die genannte Konfiguration stellt einen Appliance Server dar, der sämtliche Komponenten von Hulbee Enterprise Search enthält. Aufgrund dessen, dass hier ein vollwertiger Windows Server 2012 R2 Standard enthalten ist, kann die genannte Lösung, neben einer gesonderten Suche, auch zur Erweiterung von Diensten auf der Grundlage der MS Active Directory² und zur Speicherung von Daten in kleinen Strukturen³, die hierfür noch gleichrangige Netze nutzen, dienen.

Die Swisscows Company Search enthält Releases, die sich sowohl durch die Hardwareleistung als durch die Begrenzung der Nutzerzahlen unterscheiden.

2 Erstinstallation

2.1 Vollständige Installation

2.1.1 Indexstore

Als Indexspeicher wird Elasticsearch v.1.7 verwendet. Führen Sie die Installation entsprechend der auf der Internetseite des Herstellers enthaltenen Anleitung [siehe Kapitel 5] durch. Elasticsearch kann sowohl auf einem Computer mit Komponenten eines Processing Server als auch auf einem Einzelcomputer installiert werden. Bei der Installation von Elasticsearch auf einem Einzelcomputer kann ein anderes Betriebssystem ausser Windows, jedoch mit JRE (GNU/Linux, Solaris, etc), installiert werden.

Bei besonders hoher Auslastung wird die Nutzung eines Clusters aus mehreren Elasticsearch empfohlen.

Nach der Installation von Elasticsearch entsprechend der Anleitung sind zusätzliche Einstellungen in der Konfiguration von Elasticsearch (Datei `elasticsearch.yml`) erforderlich. Ergänzen Sie das Ende der Dateien mit folgenden strings:

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

² Links zur Einführung in Active Directory: siehe Abschnitt 5 „Hilfreiche Links“.

³ Denken Sie daran, dass ein backup installiert sein muss und regelmässig durchzuführen ist. Allgemeine Informationen über Sicherheitskopien: Siehe Abschnitt 5 „Hilfreiche Links“.

Achtung! Der Index kann vertrauliche Daten enthalten. Um Lecks solcher Daten zu verhindern, deaktivieren Sie bitte alle TCP/IP-Verbindungen für alle Komponenten ausser Application Server.

2.1.2 Verarbeitungsserver

2.1.2.1 Hardware-Voraussetzungen für HES Processing Server

Komponente	Minimum	Empfohlen
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-Voraussetzungen für HES Processing Server

Installieren Sie bitte Windows Server 2012 R2 Standard mit den neuesten Updates und den folgenden Komponenten:

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	
IIS-CommonHttpFeatures	IIS-ApplicationDevelopment	IIS-Performance	WCF-TCP-PortSharing45
IIS-Security	IIS-NetFxExtensibility45	IIS-HttpCompressionStatic	

Vor der Installation können Sie den folgenden Befehl ausführen:

```
>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-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 /All
```

Auch diese Handlung kann unter Nutzung des Systemapplets "Turn Windows features on or off" vorgenommen werden.

2.1.2.3 Domain und Benutzer-Setup

1. Verbinden Sie Server in Active Directory.
2. Erstellen Sie Benutzer **hes_user**.
3. Erteilen Sie Berechtigungen für **das Lesen** und **Ordnerinhalt auflisten** im Filestorage für **hes_user**.
4. Fügen Sie **hes_user** in die lokalen Administrator-Gruppe im HES-Verarbeitungsserver hinzu.

2.1.2.4 IFilters setup

Bitte installieren Sie folgende IFilter, um einen besseren Text-Auszug für Word- und PDF-Dateien zu bekommen:

- MS Office: <http://www.microsoft.com/en-US/download/details.aspx?id=17062> mit Service pack <http://support.microsoft.com/kb/2687447>. Installieren Sie bitte 64-Bit Version.
- PDF iFilter 64 11.0.01: <http://www.adobe.com/support/downloads/detail.jsp?ftpID=5542>.

2.1.2.5 HES Auspacken

Das HES Anwendungspaket sieht so aus: **HES.1.7.XX.XXXXX.zip** (XX ist Ihre spezifische Version). Es enthält folgende Komponenten:

- Helpers
 - ConfigTransformationHelper
- Services
 - Ses.Services.Crawler
 - Ses.Services.IndexCleaner
- Utilities
- Web
 - Ses.Web
- Ses.Setup.*

Entpacken Sie es bitte in demselben Ordner, zum Beispiel

C:\HES (i.e. C:\HES\Web\, C:\HES\Services\, C:\HES\Utilities\, etc).

2.1.2.6 Ersteinrichtung

Führen Sie bitte das Dienstprogramm ses.setup.exe im Stammordner der entpackten distributive aus.

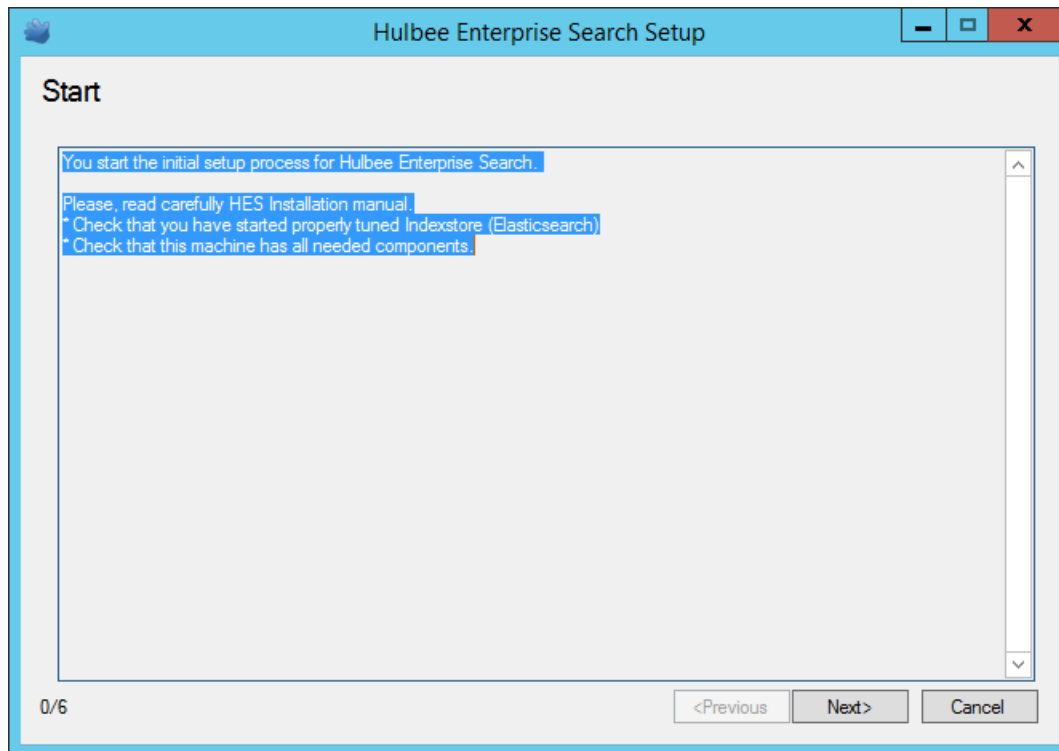


Abb. 1. HES-Installationsassistent – Beginn der Installation.

Klicken Sie bitte auf "Weiter", um zum Schritt 1 zu kommen.

1. Schritt 1.

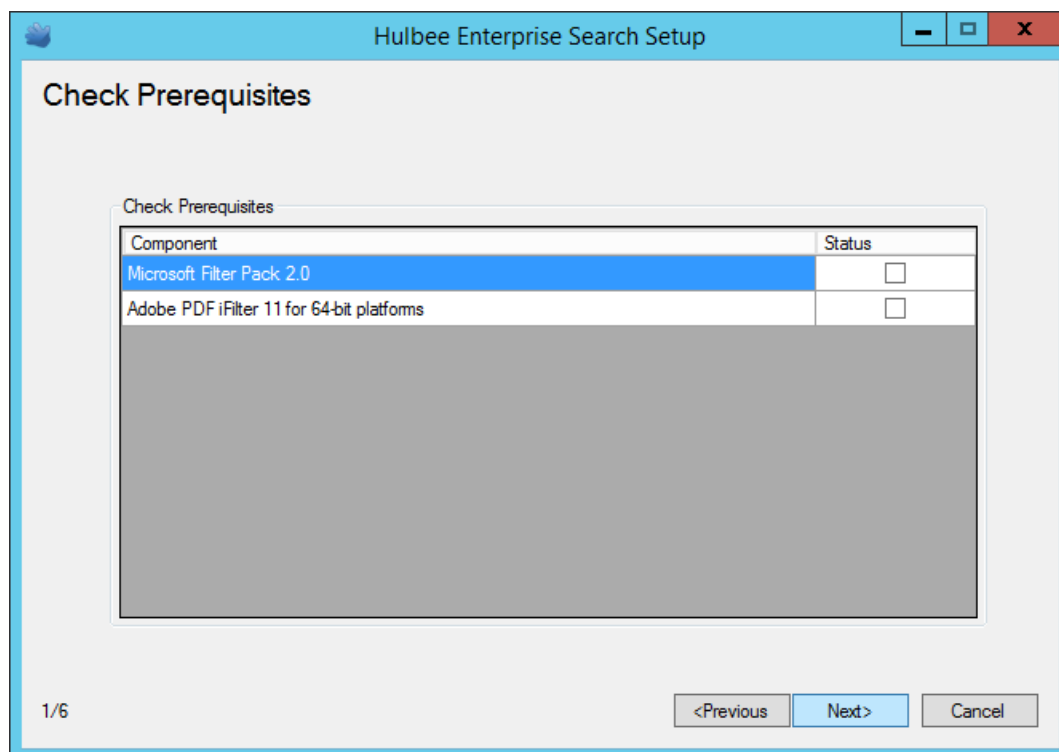
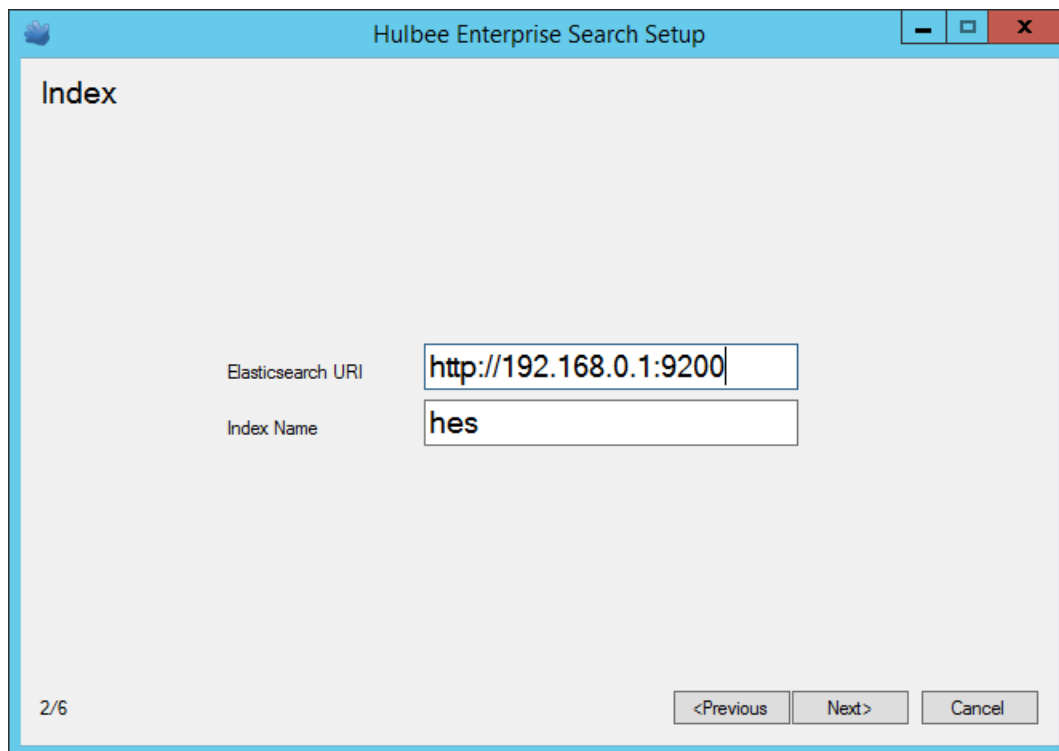


Abb. 2. Schritt 1 des HES-Installationsassistenten.

Bei der Prüfung der optionalen Voraussetzungen werden die fehlenden Komponenten angezeigt. Eine Fortsetzung der Installation ist möglich, jedoch kann durch das Fehlen des IFilters die Qualität beim Extrahieren von Text aus verschiedenen Formaten (MS Office und PDF) beeinträchtigt werden.

2. Schritt 2.

Geben Sie bitte die gültige URL zu dem Elasticsearch Server und den Indexnamen ein (Standardwert ist "hes"). URL sollte http-Protokoll-Präfix, die richtige IP-Adresse oder Domain-Namen und Port enthalten.



Hulbee Enterprise Search Setup

Index

Elasticsearch URI

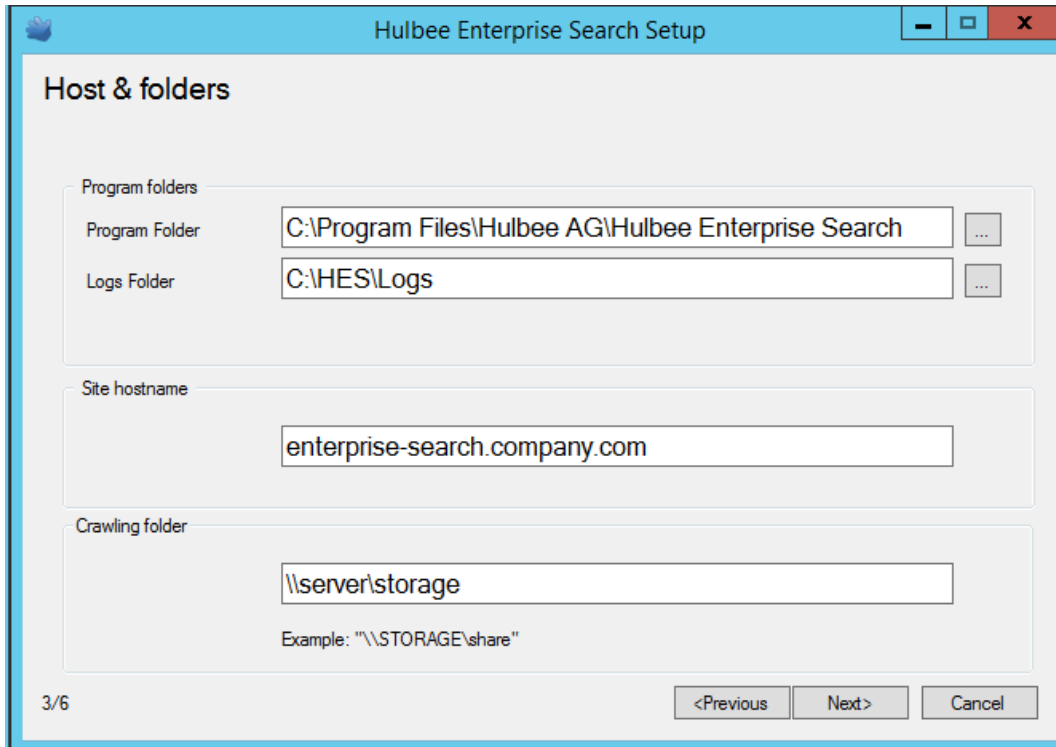
Index Name

2/6

<Previous Next> Cancel

Abb. 3. Schritt 3 des HES-Installationsassistenten.

3. Schritt 3.



The screenshot shows the 'Hulbee Enterprise Search Setup' window, specifically the 'Host & folders' step. The window has a blue title bar and standard Windows window controls. The main area is divided into three sections: 'Program folders', 'Site hostname', and 'Crawling folder'. The 'Program folders' section contains two text boxes: 'Program Folder' with the value 'C:\Program Files\Hulbee AG\Hulbee Enterprise Search' and 'Logs Folder' with the value 'C:\HES\Logs'. The 'Site hostname' section contains a text box with the value 'enterprise-search.company.com'. The 'Crawling folder' section contains a text box with the value '\\server\storage' and an example below it: 'Example: "\\STORAGE\share"'. At the bottom left, it says '3/6'. At the bottom right, there are three buttons: '<Previous', 'Next>', and 'Cancel'.

Host & folders

Program folders

Program Folder: C:\Program Files\Hulbee AG\Hulbee Enterprise Search

Logs Folder: C:\HES\Logs

Site hostname: enterprise-search.company.com

Crawling folder: \\server\storage

Example: "\\STORAGE\share"

3/6

<Previous Next> Cancel

Abb. 4. Schritt 3 des HES-Installationsassistenten.

In den meisten Fällen müssen Sie nur **"Site Hostname"** und **"Crawling Ordner"** ändern. **"Programmordner"** und **"Logs"** können mit Standardwerten belassen werden. Bitte nutzen Sie den Namen UNC-Format für **"Crawling Ordner"**.

4. Schritt 4.

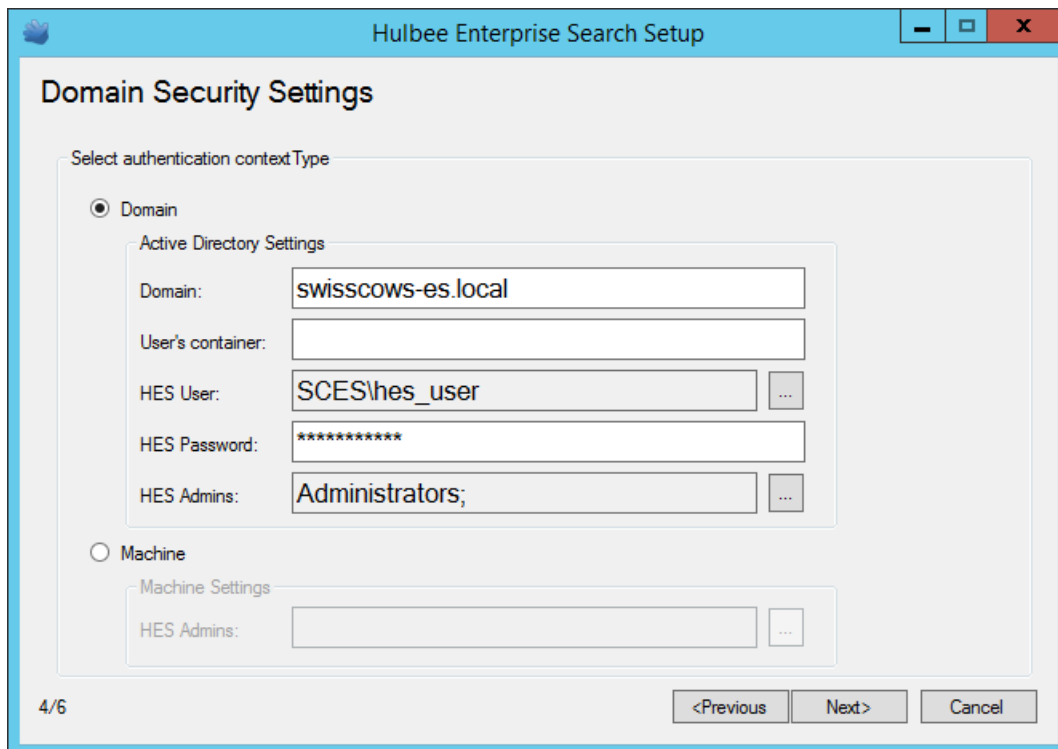


Abb. 5. Schritt 4 des HES-Installationsassistenten.

Die Felder **“Domain”**, **“HES User”** / **“Password”** entsprechen dem Benutzer "hes_user", der im Kapitel 2.1.2.3 erstellt werden soll.

Das Feld **“User’s container”** kann wie üblich ausgefüllt werden.

Für die Nuterauswahl oder die Auswahl einer Administratorgruppe klicken Sie den Button rechts vom Eingabefeld an. Hierdurch wird ein Systemdialog geöffnet.

Die Option „Machine“ ist für Testinstallationen und besondere Konfigurationen vorgesehen, die in der genannten Anleitung nicht erläutert werden.

5. Schritt 5.

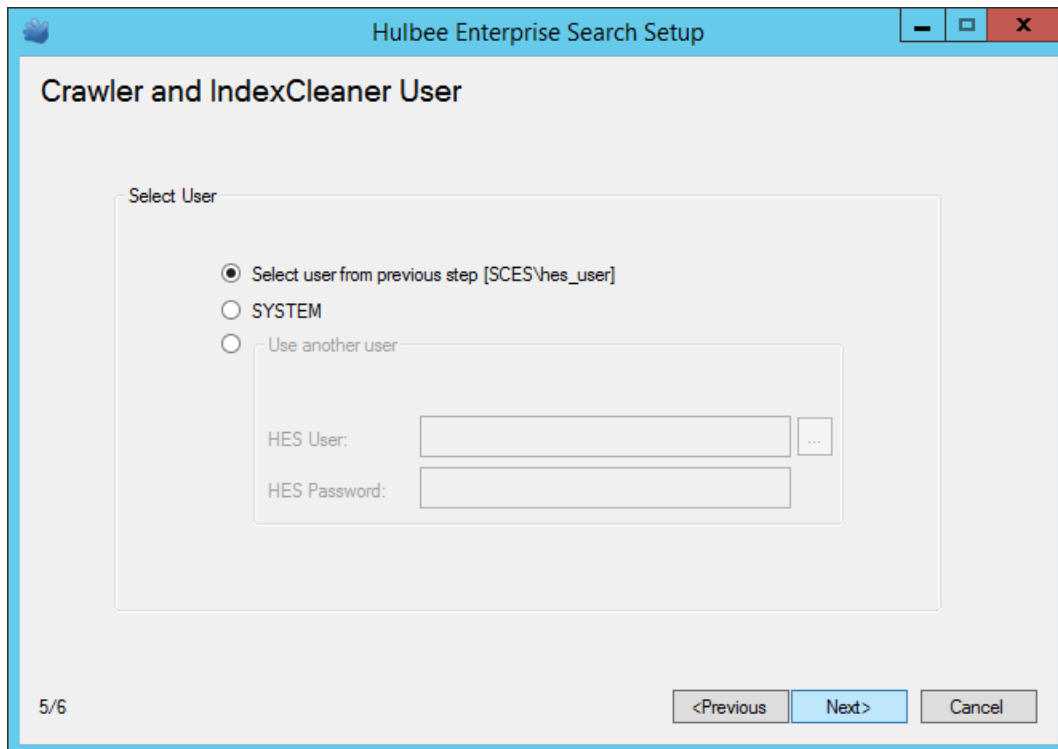


Abb. 6. Schritt 5 des HES-Installationsassistenten.

Benutzen Sie "hes_user" für die Konfiguration mit einem getrennten Filestore. Bitte vergessen Sie nicht, es in dem Verarbeitungsserver hinzuzufügen.

6. Schritt 6.

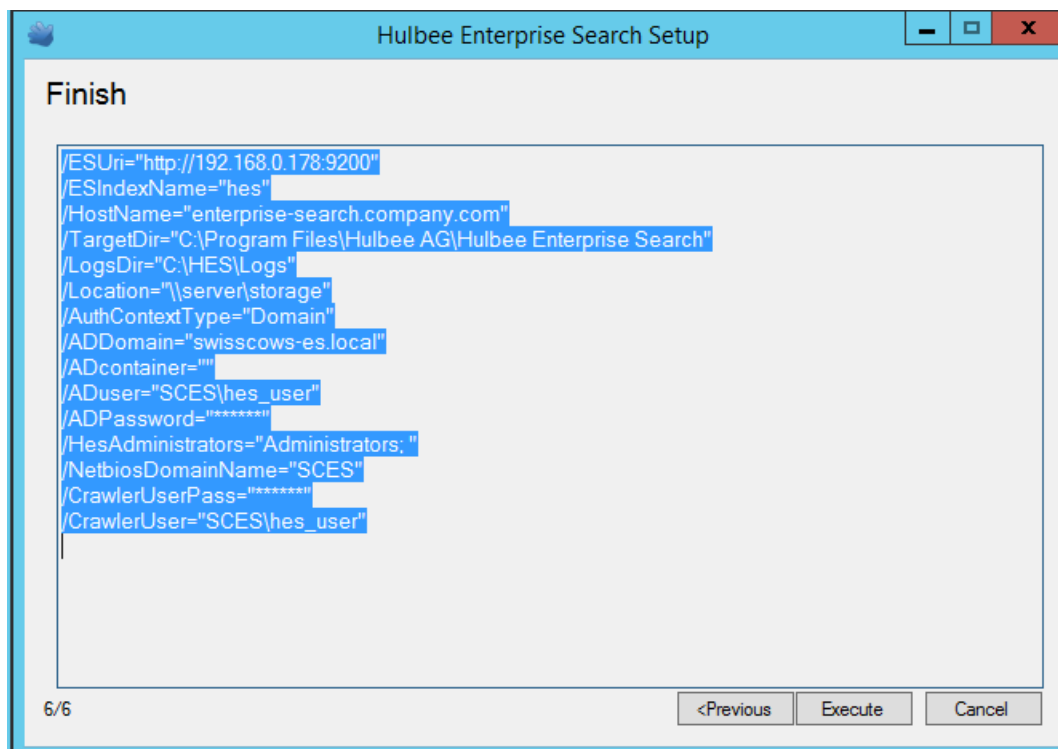


Abb. 7. Schritt 5 des HES-Installationsassistenten.

Sie können die ausgewählten Optionen ansehen und die Einstellungen mit dem “Execute” Button ausführen.

7. Letzter Schritt.

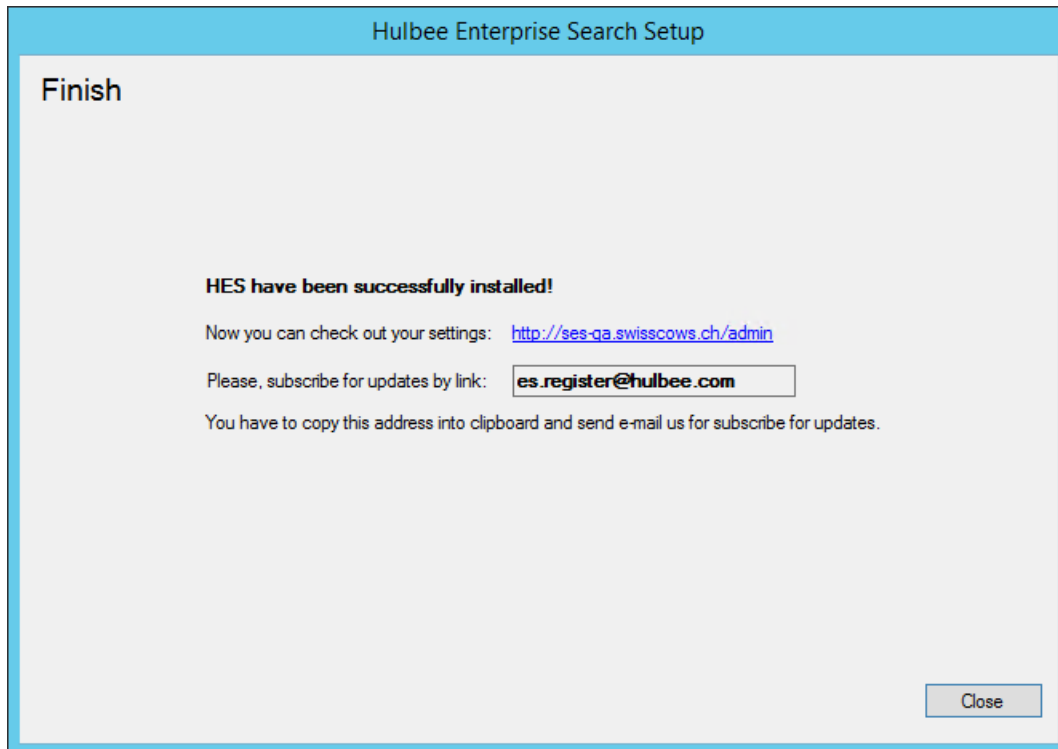


Abb. 8. HES-Installationsassistent – Abschluss der Installation.

Nach erfolgreicher Installation erscheint ein Abschlussbildschirm mit einem Link zur Administratorenleiste und mit einer Adresse. Beim Versand einer Mail mit „subscribe“ an diese Adresse kann der Produktnewsletter abonniert werden.

2.2 Swisscows Company Search

Die genannte Konfiguration erfolgt mit allen vorinstallierten Komponenten, die für die Installation von HES und für die Arbeit damit erforderlich sind. Aus diesem Grund sind solche Schritte erforderlich.

2.2.1 Einrichtung einer Domain

Planen Sie vor allem die Nutzung des bereitzustellenden Appliance Servers. In Abhängigkeit hiervon ist folgende Nutzung möglich:

- Einbindung des Swisscows Company Search Server in eine bestehende Domain.
- Einrichtung einer neuen Domain auf der Grundlage des Dienstes Microsoft Active Directory im Windows Server 2012 R2.

Einführende Informationen über den Dienst Active Directory: Links in Abschnitt 5 „Hilfreiche Links“.

Während der Bereitstellung kann sich der Administrator unter Nutzung des vorinstallierten Bedienerkontos „Administrator“ mit dem Passwort “ Admin123” auf dem Server autorisieren lassen.

Achtung! Eine Änderung des Administratorpasswortes ist unbedingt erforderlich.

Fügen Sie nach Inbetriebnahme des Servers in der Domain entsprechend der Anleitung in Abschnitt 2.1.2.3 weitere Nutzer hinzu.

2.2.2 Dateispeicher

Die übliche Grösse des Suchindex beträgt bis zu 10% der binären Grösse eines Dokuments (Richtwert, da diese Grösse im Wesentlichen von der tatsächlichen Datenauswahl abhängig ist). Der Appliance Server kann in diesem Zusammenhang auch als Netzwerkspeicher dienen.

Wenn z.B. vorauszusehen ist, dass der Dateiumfang in den nächsten Jahren nicht mehr als 1 TB beträgt, wenn jedoch der Festplattenspeicher 2 TB beträgt, kann nicht nur die Suche, sondern auch das Speichern von Dokumenten auf dem Appliance Server vollkommen sicher erfolgen. In diesem Fall wird die Belastung des Netzwerks während einer intensiven Indexierung reduziert.

Ist ein höherer Dateiumfang vorauszusehen, sind die Dateien unbedingt auf einem Server zu speichern, der mit einem für diesen Umfang ausreichenden Festplattenspeicher ausgestattet ist, oder es sind zusätzliche Festplatten im Appliance Server zu installieren.

Achtung! Der Dateispeicher muss über ein lokales Netzwerk zugänglich sein und, ebenso wie die übrigen Teile von HES (Nutzer **hes_user**), den Dienst Active Directory nutzen.

2.2.3 Einstellungen

Nach der Planung der Nutzung des Appliance Servers im Bereich des Dienstes Active Directory und nach der Anordnung des Dateispeichers sind unbedingt die Installationsschritte gemäss 2.1.2.6 Ersteinrichtung durchzuführen.

3 Überprüfung der Installation

Um mit der Suche zu beginnen, geben Sie den URL im Browser ein. Dieser wurde im Kapitel 2.1.2.6, Schritt 3, definiert:

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

Falls das Benutzerkonto nicht automatisch erkannt wurde, geben Sie Benutzer (inkl. NetBIOS Domainname) und Passwort ein.

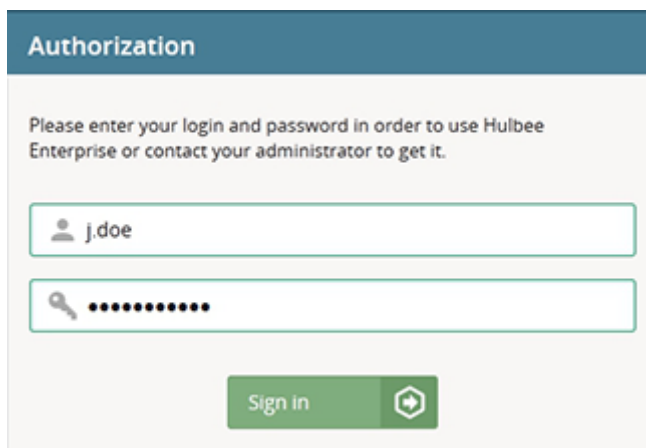


Abb. 9. Formen der Autorisierung.

Nach der Anmeldung erscheint die Startseite:

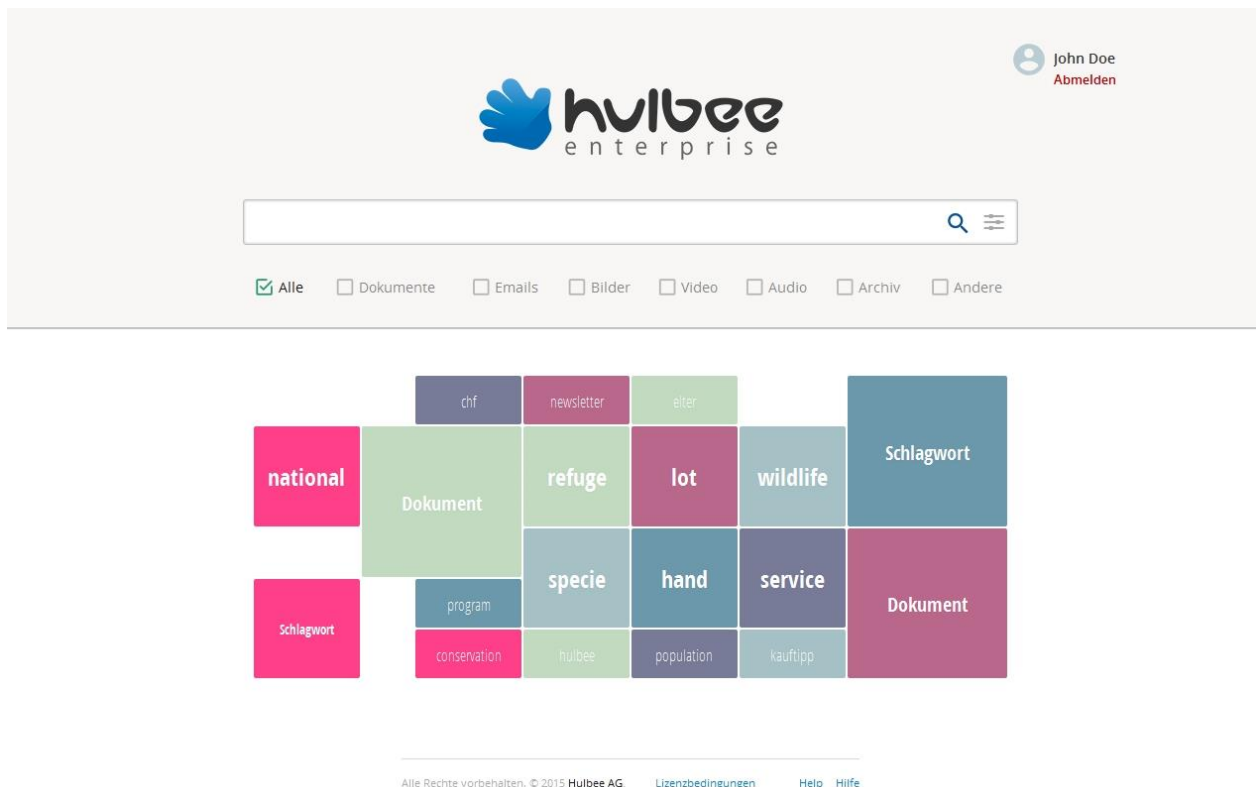


Abb. 10. HES-Startseite.

Sie müssen eine Suchanfrage eingeben, um Dokumente mit den relevanten Informationen zu erhalten.

4 Einrichtung der Admin Seite

Die übrigen Setup-Schritte können im Admin-Panel durchgeführt werden. Bitte folgen Sie dem nächsten Link:

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

Geben Sie anstelle von "enterprise-search.company.com" den Link ein, der für den Installationsschritt 3 genutzt wurde und in Abb. 4 angegeben ist.

4.1 Wiederherstellungsoptionen

Wenn Sie bereits ein abgestimmtes Exemplar von HES installiert haben, können Sie die zuvor gesicherten Einstellungen wiederherstellen. Bitte, wählen Sie "Export / Import" und:

1. Klicken Sie auf "Datei auswählen" und wählen Sie die Sicherungsdatei aus.
2. Klicken Sie auf "Importieren".

Dieser Vorgang ermöglicht die Wiederherstellung globaler und persönlicher Benutzer-Einstellungen. Das heisst, dass es sich um die Daten handelt, die zu sehen sind und mithilfe der Administratorleiste und

mithilfe des Userboards bearbeitet werden. Der Suchindex ist nicht Bestandteil des Einrichtens – er wird mithilfe des Crawlings erneut aktualisiert.

4.2 Administrationsbereich

Die Administratorleiste besteht aus einem Navigations- und einem Arbeitsbereich. Der Navigationsbereich enthält Links, die einen Übergang in den einen oder anderen Unterbereich der Administratorleiste ermöglichen. Für alle Seiten gibt es nur eine einzige Administratorenleiste. Der Arbeitsbereich enthält Elemente für die Bearbeitung von Einstellungen und für die Anzeige von verschiedenen Informationen über den Betrieb des Systems.

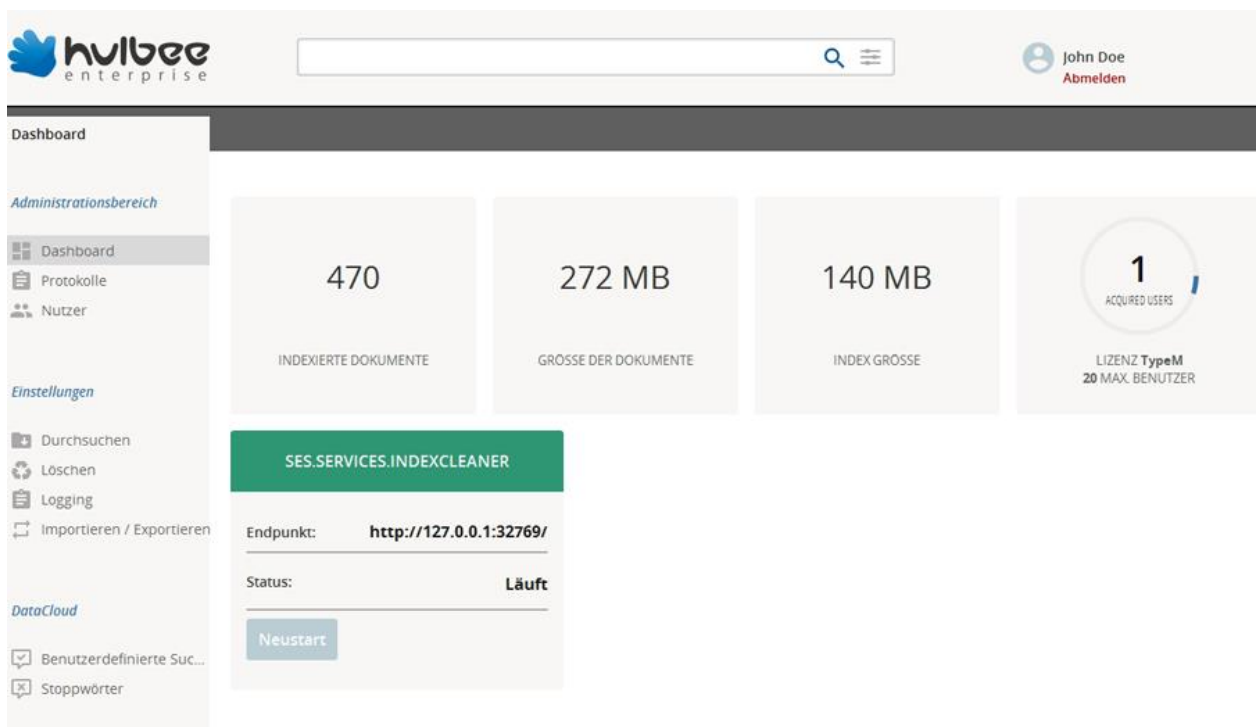


Abb. 11. Administratorsite. Allgemeine Ansicht.

In den Screenshots werden lediglich die Bereiche dargestellt, die beim Wechsel zwischen den unterschiedlichen Links der Administratorenleiste geändert werden.

4.2.1 Dashboard

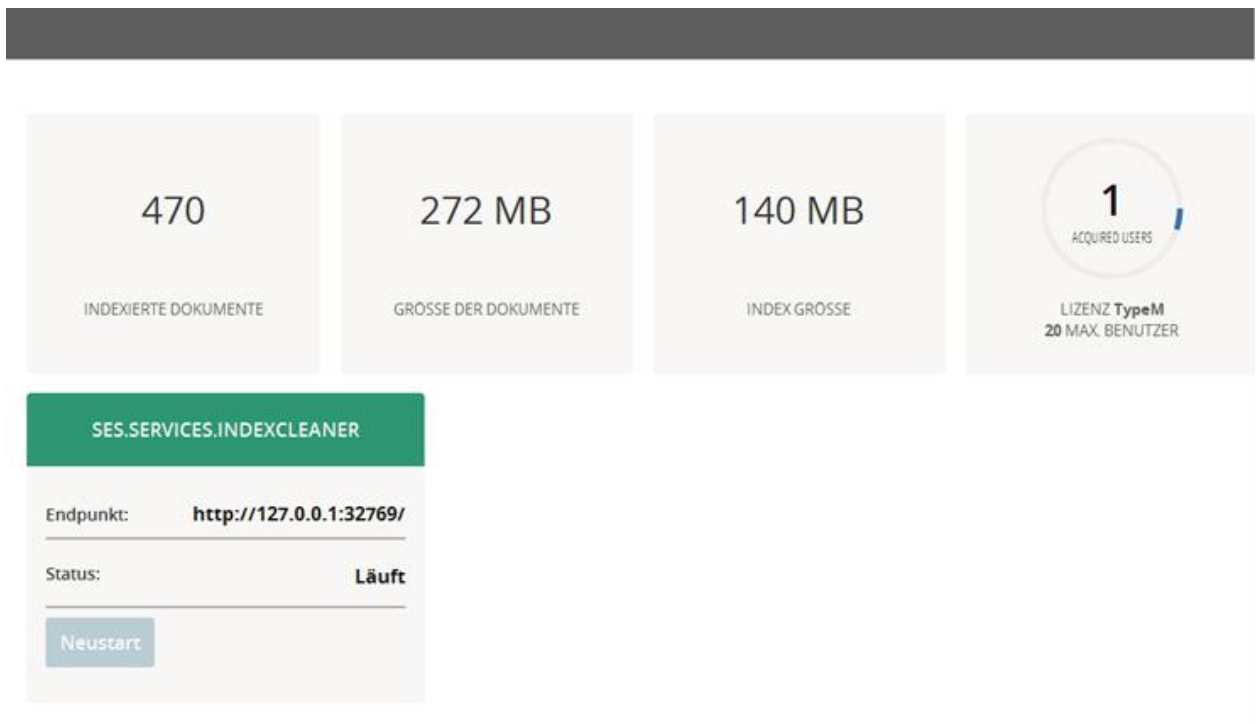


Abb. 12. Dashboard.

Auf dem Dashboard sind verschiedene Statistiken dargestellt:

- INDEXED DOCUMENTS. Anzahl der indexierten Dokumente. Diese Zahl kann nicht mit der Gesamtzahl der im Speicher vorhandenen Dateien übereinstimmen, da ein Teil der Dokumente aufgrund verschiedener Filter (Grösse, Erweiterungen) nicht bearbeitet werden kann. Gleichzeitig kann in den Archiven und Mitteilungen mehr als eine Datei enthalten sein.
- DOCUMENTS SIZE. Gesamtgrösse der im Index vorhandenen Dokumente. Hier ist die Summe der Binärgrössen gemeint. Auch dieser Wert weicht bei dem von der Datei benötigten Speicherplatz ebenfalls aus den vorgenannten Gründen ab.
- INDEX SIZE. Indexgrösse. Elasticsearch auf der Festplatte.
- LICENSE {id}. In diesem Widget ist der Lizenztyp angegeben. Anzahl der Nutzer, die den Server bereits genutzt haben und Anzahl der potenziellen Nutzer.
- Servicewidgets. Mithilfe dieser Widgets kann man sich vom einwandfreien Crawling und vom einwandfreien Indexcleaning überzeugen. Sie ermöglichen ausserdem einen Neustart nach Änderung der Einstellungen.

4.2.2 Logs

Seite 1 von 37		Veranstaltungen: Kritisch, Fehler, Warnung, Information, Wortreich, Start, Stopp	
ZEITSTEMPEL	VERANSTALTUNG	MELDUNG	PROZESS
Heute, 10:48:38	Information	Deleting old logs	4456Ses.Services.IndexCleaner6
Heute, 10:48:34	Information	Application started at 12/3/2015 10:48:34 AM	4456Ses.Services.IndexCleaner....
Heute, 10:48:17	Information	Application started at 12/3/2015 10:48:16 AM	4744Ses.Services.Crawler.exe1
Gestern, 11:43:41	Error	Error occured while processing file "\\localhost\SHARE\daten\Calibre-Bibliothek\Hans Christian Andersen\Fairy Tales of Hans Christian Anders (50)\Fairy Tales of Hans Christian A - Hans Christian Andersen.pdf": System.AggregateException: One or more errors occurred. --> System.ApplicationException: Worker don't send reply. at	5980Ses.Services.Crawler14

Abb. 13. Logs.

Auf dieser Seite können Fehlermeldungen, Warnungen und andere Informationen geprüft werden. Eine Filterung der Ereignisse kann durch Setzen der entsprechenden Häkchen in der sich öffnenden Liste „Events“ vorgenommen werden.

4.2.3 Benutzer

Seite 1 von 1	
Benutzerdefinierte Suchanfragen	
<input type="checkbox"/>	BENUTZERDEFINIERTE SUCHANFRAGE
<input type="checkbox"/>	Dokument
<input type="checkbox"/>	Schlagwort
*	<input type="text" value="Geben Sie eine neue Suchanfrage ein..."/>

Abb. 14. Benutzer

In der Benutzerseite kann man sehen, wer als Nutzer im System bereits autorisiert ist. Hier kann für einen Nutzer die HES-Suche HES auch verweigert werden. In diesem Fall bleibt der betreffende Nutzer bei der Prüfung der Richtigkeit der Anzahl der Lizenznutzer unberücksichtigt. Bei der Meldung «Zugriff verweigert» bleibt dem Nutzer und Administrator nur der Zugang zur Administratorenleiste.

4.3 Einstellungen

4.3.1 Crawling

Suchvorgang Orte
Um eine oder mehrere Standorte auf dem lokalen Computer oder im lokalen Netzwerk zu durchsuchen, müssen diese durch Semikolon getrennt werden.

Suchvorgang Platzhalter
Der Suchmuster muss mit den Dateinamen im Suchvorgang übereinstimmen.

Ignorieren
Ein Filter mit dem Dateien oder Verzeichnisse beim Durchsuchen ausgeschlossen werden können. (Leer lassen, um alle Dateien und Verzeichnisse zu durchsuchen.)

Anzahl der Parallelarbeiten
Die maximale Anzahl von Parallelarbeiten, die beim Crawling-Vorgang ausgeführt werden können.

Zeitüberschreitung für das Senden der Antwort durch den Crawler

Zeitüberschreitung für das Senden der Aufgabe durch den Crawler

Abb. 15. Crawling.

Bitte wählen Sie "Crawling" und passen Sie die folgenden Felder an:

- "Scan/watch locations": Überprüfen Sie und ändern Sie bei Bedarf die Standorte auf dem lokalen Computer oder im lokalen Netzwerk, die gescannt und überwacht werden sollen, durch ein Semikolon. Dieser Pfad sollte in der UNC-Form (zB "\\ DC \ Files") sein.
- "Scan/watch files wildcard": Baustein für die zu bearbeitenden Dokumente. Empfohlene Adresse "*. *". Es können auch mehrere Bausteine, aufgezählt mit Komma, genannt werden. Diese wird gewöhnlich für Tests genutzt und bezieht sich lediglich auf den Dateinamen.
- "Ignore patterns": Baustein, durch den bestimmte Dateien oder Ordner unberücksichtigt bleiben können. Hier ist ein vollständiger Pfad anzugeben. "*.log" – hier bleiben alle Dateien mit der Endung ".log" unberücksichtigt. "*\fo*" – hier bleiben alle Dateien, deren Namen mit "fo" beginnen, unberücksichtigt. Als einziger Platzhalter dient das Symbol "*". Es können auch mehrere Bausteine, aufgezählt mit Komma, genannt werden.
- "Concurrent workers count": Um eine bessere Verarbeitungsgeschwindigkeit zu erhalten, können Sie die Anzahl der CPU-Kerne eingeben. Für die Konfiguration des Elasticsearch, das sich auf einem anderen Server befindet, wird die Festlegung einer Anzahl von Dateimanagern, analog zur Anzahl der Kerne des Computers mit Elasticsearch, empfohlen. Jedoch darf diese nicht höher als die Anzahl der Kerne des Processing Server sein. Für die Konfiguration von Swisscows Company Search sowie in den Fällen, in denen alle Module in einem Computer

vereint sind, wird die Installation von Dateimanagern empfohlen, deren Anzahl nur die Hälfte der Anzahl der CPU-Kerne betragen darf.

- “Timeout for send reply by worker”: Zeit (ms), die dem Dateimanager für die von Befehlen (eines Dokuments) zugewiesen wird.
- “Timeout for send reply by crawler”: Wartezeit (ms) des Dateimanagers auf einen Befehl. Bleiben die Befehle aus, schaltet sich der Dateimanager ab, um die Ressourcen zu schonen.

Für den ersten Zyklus der Indexierung ist es sinnvoll, eine kurze Zeitspanne, ca. 15 Sekunden (15000 ms), festzulegen, um eine schnellere Traversierung aller Dokumente zu ermöglichen. Nach der ersten Traversierung kann die Zeitspanne für das Timeout (z.B. auf 120000 ms) verlängert werden, um auch eine Traversierung von übergrossen Dateien (Archive, grosse Dokumente), die im ersten Zyklus nicht verarbeitet werden konnten, zu ermöglichen.

☐ Ein benutzerdefinierter Elasticsearch Server benutzen

Uri vom Elasticsearch Server

Standard Index Name

Die maximale Größe des Anhangs

Die maximale extrahierte Größe








Der maximal zu verarbeiteten Grenzwert

Abb. 16. Crawling. Fortsetzung.



- “Use custom Elasticsearch server”: Diese Konfigurationsoption kann nicht zugeordnet werden.
- “Maximum attachment size”: Maximale Grösse der abgelegten Dateien. Wenn diese erreicht ist, wird die Verarbeitung des Archivs oder von Mitteilungen abgebrochen.
- “Maximum extracted size”: Maximale Grösse eines Textes in Symbolen, der zur weiteren Verarbeitung und Indexierung vorgelegt wird. Das bezieht sich speziell auf einen Text, der aus einem Dokument extrahiert wurde, und nicht auf die binäre Dateigrösse. Der übrige Text wird abgetrennt.
- “Maximum processed common limit”: Maximale Grösse einer Datei, deren Konvertierung versucht wurde. Bei grossen Dateien bleibt lediglich die Metainformation (Name, Pfad, Tag der Erstellung und Tag der Änderung, Grösse) erhalten.
- “Filters by extensions”: Einrichtung eines Konvertierungsprogramms in Abhängigkeit von der Dateierweiterung. IFilter arbeiten in der Regel langsam, aber sie erzielen bei Officedokumenten und Dokumenten im PDF-Format die besten Ergebnisse.
- “Maximum processed limits by extensions”: Verschiedene Dateitypen haben ein grosses Volumen und überschreiten die allgemein als Grenzwert festgelegte Grösse. Jedoch ist eine

Verarbeitung solcher Dateien einfach, weil hier nur ein kleiner Teil der Daten verarbeitet wird. Im Allgemeinen handelt es sich hierbei um Mediadateien der Formate "avi; wmv; mp4; mkv". Hierfür kann das Volumen der genannten Dateitypen individuell begrenzt werden. Das kann durch einen String erfolgen, der durch das Symbol ";" erweitert wird.

Filter anhand der Dateierendungen

pdf	ifilter dann nativev	
docx;doc;docm	ifilter dann nativev	
ppt; pptx	ifilter dann nativev	
xls;xlsx	ifilter dann nativev	
one	ifilter dann nativev	
csv	nur native	
<input type="text"/>	ifilter dann nativev	

Die maximal zu verarbeiteten Grenzwerten anhand der Dateierendungen

avi;wmv;mp4;mkv	8589934592	
<input type="text"/>	1	


Einstellungen speichern 

Abb. 17. Crawling. Ende.

Nach Änderung dieser Optionen, klicken Sie bitte auf "Einstellungen speichern". Die neuen Einstellungen können durch das Crawling nach dem Neustart genutzt werden. Das kann mithilfe des entsprechenden Dashboardwidgeets erfolgen. Ist jedoch aus irgendwelchen Gründen kein Zugriff von der Administratorenleiste aus möglich (bei fehlendem Widget), können hier auch die System Applet Services genutzt werden.

4.3.2 Cleaning

Losgröße

Die Anzahl der Dokumente, die pro eine Iteration verarbeitet werden.

Iteration Zeitüberschreitung

Die Zeitüberschreitung zwischen den Iterationen (in Format "hh:mm:ss" 00:05:00 - fünf Minuten).

☐ Ein benutzerdefinierter Elasticsearch Server benutzen

Uri vom Elasticsearch Server

Standard Index Name

Einstellungen speichern



Abb. 18. Cleaning.

Bitte wählen Sie "Cleaning" und passen Sie die folgenden Felder an:

- "Batch size": Anzahl der Dokumente, deren Aktualität in einem einzigen Schritt geprüft werden kann. Es wird der Wert 1000 empfohlen.
- "Iteration timeout": Pause bei der Verarbeitung der einzelnen Dateiteile. In den meisten Fällen beträgt diese Pause 10 Sekunden. Hierbei handelt es sich um eine angemessene Zeitspanne. Geht es jedoch um die Verarbeitung von Millionen oder zehn Millionen Einheiten, kann diese Zeitspanne verringert werden, um eine schnellere Löschung von Dokumenten, die aus dem Index gelöscht wurden (oder deren Speicherpfad geändert wurde), zu erreichen (jedoch wird die Belastung des Processing Servers und des Indexstores erhöht).

Nach Änderung dieser Optionen, klicken Sie bitte auf "Einstellungen speichern". Die neuen Einstellungen können vom IndexCleaner nach dem Neustart genutzt werden. Das kann mithilfe des entsprechenden Dashboardwidgets erfolgen. Ist jedoch aus irgendwelchen Gründen kein Zugriff von der Administratorenleiste aus möglich (bei fehlendem Widget), können hier auch die System Applet Services genutzt werden.

4.3.3 Logging

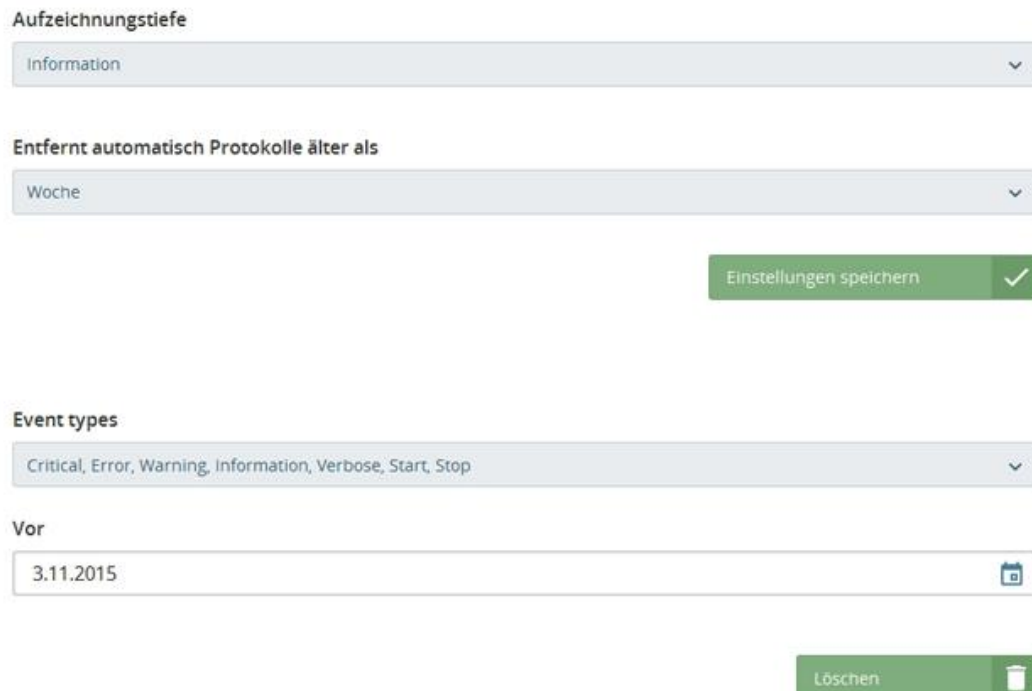


Abb. 19. Login.

Die Seite für die Login-Einstellungen ermöglicht folgende Einstellungen:

“Trace level”: Hier öffnet sich eine Liste, in der eingestellt werden kann, welche Ereignisse im Systemjournal aufgezeichnet werden sollen.

“Automatically remove logs older than”: Hier kann die Anzahl der gespeicherten HES-Logins festgelegt werden. Das Cleaning erfolgt sowohl im IndexStore (diese werden in den Administratoreinstellungen angegeben), als auch in Form von Textdateien auf einem Computer mit Processing Server (üblicher Speicherort: c:\hes\logs). Im Laufe der Zeit kann hier ein grosses Volumen angehäuft werden. Deshalb wird empfohlen, die genannte Option nicht abzuschalten, ausser in bestimmten Fällen. Das Cleaning der Logs erfolgt durch den IndexCleaner. Wird der Speicherzeitraum geändert, ist ein Neustart des Programms erforderlich.

Auf dieser Site kann auch ein manuelles Cleaning der im IndexStore gespeicherten Ereignisse erfolgen. Hierfür sind die Art der für das Cleaning vorgesehenen Ereignisse und der Tag, bis zu dem das Cleaning zu erfolgen hat, festzulegen. Klicken Sie danach den Button „Delete“ an.

4.3.4 Import / Export



Abb. 20. Import / Export.

Die Import / Export-Site enthält Befehle, die eine Speicherung der HES-Einstellungen in einer Datei und die Wiederherstellung derselben aus der Datei heraus ermöglichen. Die HES-Einstellungen können sowohl als Administratoreinstellungen als auch als Einstellungen, die vom Nutzer im persönlichen Bereich eingegeben werden, gespeichert werden. Die genannten Befehle können bei Abstürzen oder bei der Installation von Updates angewendet werden.

4.4 Datacloud

Diese Einstellungen verhalten sich analog zu den Einstellungen im persönlichen Bereich des Nutzers, jedoch beziehen sie sich nicht auf einen einzelnen HES-Nutzer, sondern auf alle gleichzeitig.

4.4.1 Benutzerdefinierte Abfragen

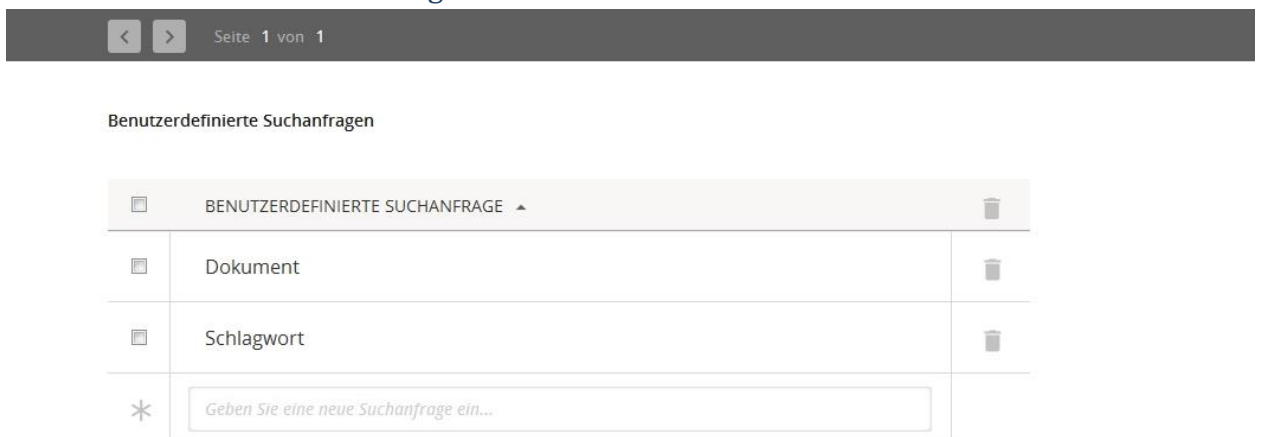


Abb. 21. Benutzerdefinierte Abfragen.

Die Einstellung « Benutzerdefinierte Abfragen» ermöglicht das Ergänzen von Schlüsselwörtern, die danach in der DataCloud auf der Startseite aller Nutzer dargestellt werden und eine schnellere Eingabe von üblichen Suchanfragen ermöglichen.

Die Nutzeranfragen werden mithilfe der Tastatur in die Eingabezeile im Bereich rechts eingegeben. Nach der Eingabe wird der Button <Enter> angeklickt. Bei Bedarf kann eine weitere Nutzeranfrage eingegeben werden. Die Nutzeranfragen können erweitert, bearbeitet oder gelöscht werden. Es ist auch eine

Navigation auf der Anfragenliste, speziell mit einzelnen Arbeitsblättern (wenn die Liste mehr als eine Seite umfasst), möglich.

4.4.2 Stoppwörter



Abb. 22. Stoppwörter.

In der Registerkarte «Stoppwörter» kann der Administrator Begriffe angeben, die in der DataCloud, die sich in der Liste mit den gefundenen Begriffen befindet, nicht angezeigt werden sollen, z. B. der Name des Unternehmens des Nutzers, der praktisch in jedem Dokument angegeben sein kann, der aber diesbezüglich für die Präzisierung der Anfrage keinen Nutzen bringt.

Die Arbeit mit der Stoppwörter-Liste erfolgt parallel zur Arbeit mit der Anfragenliste in der Registerkarte «Benutzerdefinierte Abfragen».

5 Hilfreiche Links

Active directory:

- <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

6 Bekannte Probleme

- Der Logs-Ordner sollte "Volle Kontrolle" - Rechte für alle Benutzer haben. Diese Rechte wurden bereits für den Standardordner in Swisscows für Office-Server gewährt, aber Sie müssen sie manuell hinzufügen, wenn Sie das Standortprotokoll ändern möchten.
- HES Deinstallieren. Im Falle einer erfolglosen Deinstallation von HES (HES Dienste laufen immer noch im Dienste-Applet, im Programme und Funktionen Applet ist immer noch eine Verknüpfung vorhanden), versuchen Sie eine manuelle Deinstallation:
 1. Stoppen Sie die Dienste Ses.Services.Crawler und Ses.Services.IndexCleaner mithilfe des "Dienste" Applets.
 2. Starten Sie die Befehlskonsole (cmd.exe) als Administrator.
 3. Löschen Sie die Dienste mit den folgenden Befehlen:
 - sc Ses.Services.Crawler löschen
 - sc Ses.Services.IndexCleaner löschen
 4. Wenn die Dienste noch unter "Dienste" sichtbar sind, starten Sie den Server neu.
 5. Löschen Sie den Ordner mit dem installierten HES. Es ist c:\Programme\Hulbee AG\Hulbee Enterprise Search\Standardordner.
 6. Öffnen Sie "Programme und Funktionen", und löschen Sie die "Hulbee Enterprise Search" Verknüpfung. Klicken Sie auf "Deinstallieren" und die Option „Verknüpfung löschen“ wird vorgeschlagen werden.
- IIS_IUSRS Gruppe sollte Zugriff zum Lesen der Dateien in den durchforsteten Speichern haben, um die Vorschau für Mediendokumente (Audio, Video, Bilder) zu ermöglichen.