How to Block Content Type Based Character Sets

Contents

Introduction

This document describes how to write and configure a filter in order to detect and take action on content type based character sets on the Cisco Email Security Appliance (ESA). The following document can be used to detect foreign language based characters seen in spam messages.

Background Information

ESA administrators may receive an influx of mail messages that contain character based foreign languages that are not legitimate mail for their company or domain(s). One way to address from the ESA, we have three options:

3. Write a filter using condition Message Language. (This option is a new feature for AsyncOS Email Security 10.0.0-203 and newer.)

How to Block Content Type Based Character Sets

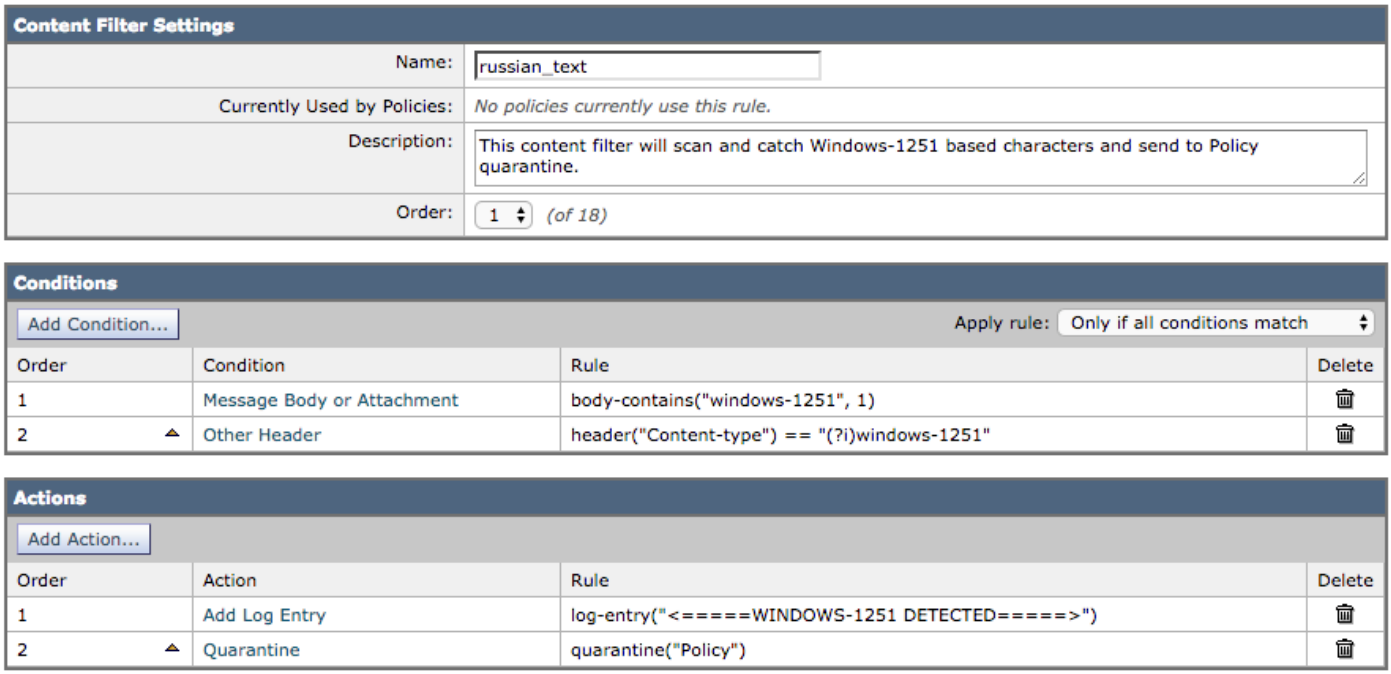
Write a filter to detect content type

The first option is for the administrator to write and configure a filter, and associate it to a mail policy, as needed.

> **Note**: Writing and configuring this filter as a message filter may be resource-expensive in order to scan the body of emails for the character sets.

> **Note**: Configuring this as a content filter is strongly suggested, as content filters occur after anti-spam scanning. However, this can be written and configured as a message filter, if needed.

The following example will take into account a mail message contain Russian (Cyrillic) based characters via the Windows-1251 based character set. Written as a content filter:



Content Filter Settings

Name:	russian_text
Currently Used by Policies:	*No policies currently use this rule.*
Description:	This content filter will scan and catch Windows-1251 based characters and send to Policy quarantine.
Order:	1 *(of 18)*

Conditions

Add Condition... Apply rule: Only if all conditions match

Order	Condition	Rule	Delete
1	Message Body or Attachment	body-contains("windows-1251", 1)	
2	Other Header	header("Content-type") == "(?i)windows-1251"	

Actions

Add Action...

Order	Action	Rule	Delete
1	Add Log Entry	log-entry("<=====WINDOWS-1251 DETECTED=====>")	
2	Quarantine	quarantine("Policy")	

The test email used will contain the following in the body of the email:

```
Russian uses , , , , o, , , , ,  as vowels. You could create a message filter set to "Matches
any of the following" that test whether "Body" "contains" "", "Body" "contains" "" and so forth
until you covered all of the vowels. Ssince English also uses "a" , "e" , "o", and "y" letters
don't test for them. The reason for "Matches any of the following" is to logically OR them - you
want the action to take place if any of those letters are found.
```

With the content filter configured as above, the mail logs would record similar to the following:

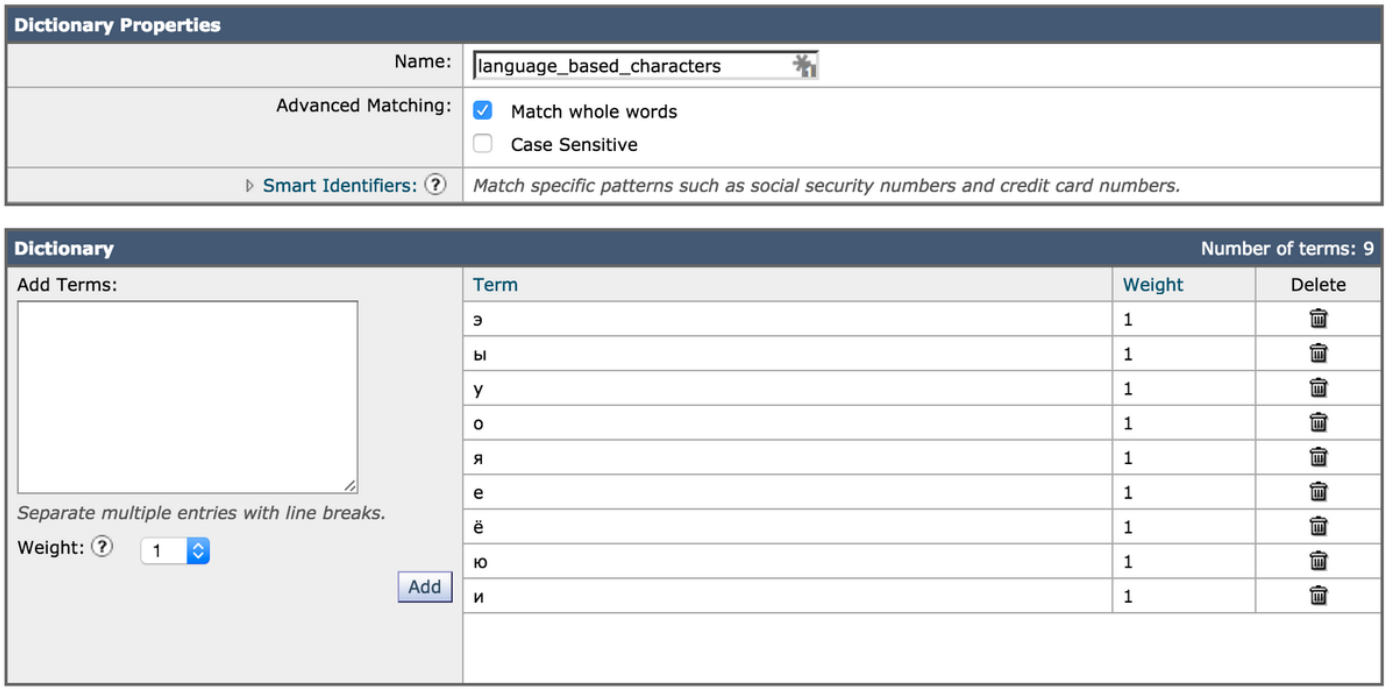
```
Thu Sep 10 14:50:09 2015 Info: Start MID 164993 ICID 266729
Thu Sep 10 14:50:09 2015 Info: MID 164993 ICID 266729 From: <end_user@test.com>
Thu Sep 10 14:50:09 2015 Info: MID 164993 ICID 266729 RID 0 To: <recpient@my_co.com>
Thu Sep 10 14:50:09 2015 Info: MID 164993 using engine: SPF Verdict Cache using cached verdict
Thu Sep 10 14:50:09 2015 Info: MID 164993 Message-ID '<7A961F85-A5F1-413F-87CB-
C31D2E5605EC@my_co.com>'
Thu Sep 10 14:50:09 2015 Info: MID 164993 Subject 'russian test'
Thu Sep 10 14:50:09 2015 Info: MID 164993 ready 2302 bytes from <end_user@test.com>
Thu Sep 10 14:50:09 2015 Info: MID 164993 matched all recipients for per-recipient policy
DEFAULT in the inbound table
Thu Sep 10 14:50:09 2015 Info: MID 164993 AMP file reputation verdict : CLEAN
Thu Sep 10 14:50:09 2015 Info: MID 164993 using engine: GRAYMAIL negative
Thu Sep 10 14:50:09 2015 Info: MID 164993 Custom Log Entry: <====== WINDOWS-1251 DETECTED
======>
Thu Sep 10 14:50:09 2015 Info: MID 164993 quarantined to "Policy" (content filter:russian_text)
Thu Sep 10 14:50:09 2015 Info: Message finished MID 164993 done
```

Other languages and character sets can be used. Please see the References section for additional information.

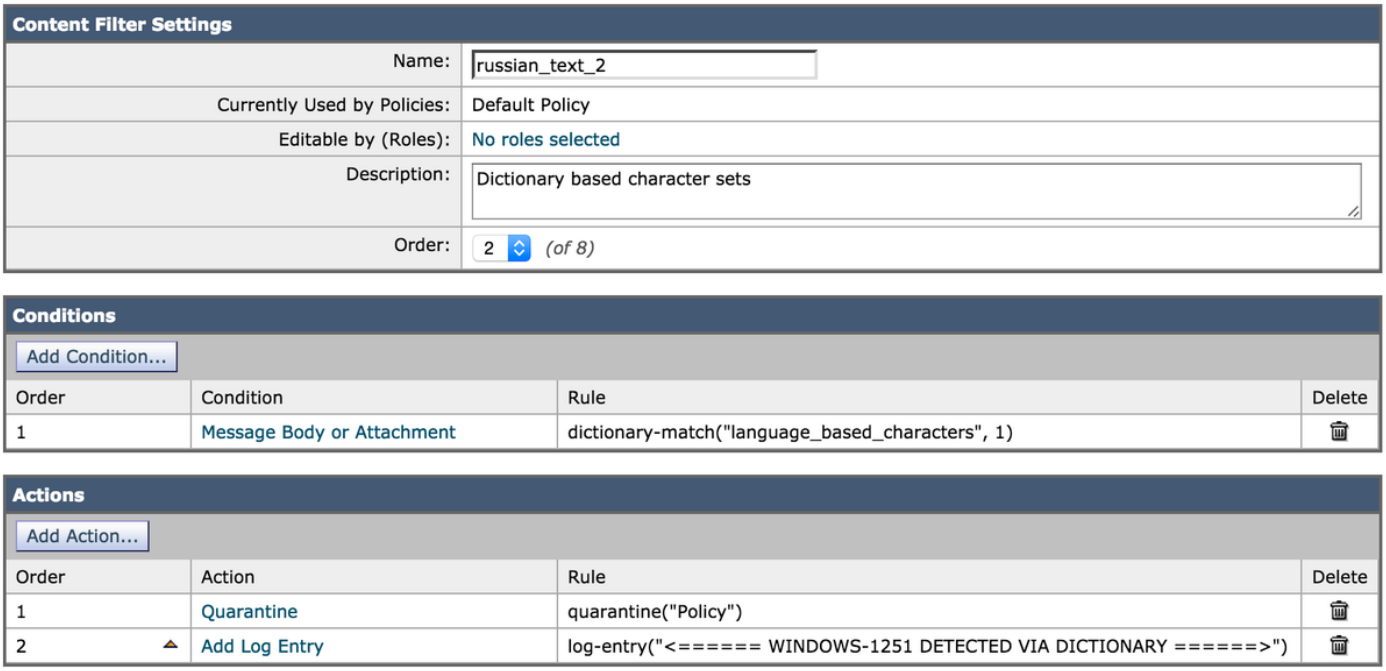
Write a filter to reference a character based dictionary

The second option is to add the list of character sets to a dictionary text file and refer to that in the filter.

Example of adding the characters to the dictionary:



The characters are now assigned to the dictionary and the dictionary itself is referenced in the condition items for the filter:



Using the same test email as above, it contains the following in the body of the email:

```
Russian uses , , , , o, , , , ,  as vowels. You could create a message filter set to "Matches
any of the following" that test whether "Body" "contains" "", "Body" "contains" "" and so forth
until you covered all of the vowels. Ssince English also uses "a" , "e" , "o", and "y" letters
don't test for them. The reason for "Matches any of the following" is to logically OR them - you
want the action to take place if any of those letters are found.
```

With the content filter configured as above using the dictionary match condition, the mail logs would record similar to the following:

```
Thu Sep 10 15:26:08 2015 Info: Start MID 164995 ICID 266737
Thu Sep 10 15:26:08 2015 Info: MID 164995 ICID 266737 From: <end_user@test.com>
Thu Sep 10 15:26:08 2015 Info: MID 164995 ICID 266737 RID 0 To: <recpient@my_co.com>
Thu Sep 10 15:26:08 2015 Info: MID 164995 using engine: SPF Verdict Cache using cached verdict
```

```
Thu Sep 10 15:26:08 2015 Info: SPF Verdict Cache cache status: hits = 6, misses = 4, expires =
1, adds = 4, seconds saved = 0.50, total seconds = 0.85
Thu Sep 10 15:26:08 2015 Info: MID 164995 Message-ID '<BCC88307-EB91-476E-8732-
334E9EE84EC8@my_co.com>'
Thu Sep 10 15:26:08 2015 Info: MID 164995 Subject 'russian test 3'
Thu Sep 10 15:26:08 2015 Info: MID 164995 ready 2316 bytes from <end_user@test.com>
Thu Sep 10 15:26:08 2015 Info: MID 164995 matched all recipients for per-recipient policy
DEFAULT in the inbound table
Thu Sep 10 15:26:08 2015 Info: MID 164995 AMP file reputation verdict : CLEAN
Thu Sep 10 15:26:08 2015 Info: MID 164995 using engine: GRAYMAIL negative
Thu Sep 10 15:26:08 2015 Info: MID 164995 Custom Log Entry: <====== WINDOWS-1251 DETECTED VIA
DICTIONARY ======>
Thu Sep 10 15:26:08 2015 Info: MID 164995 quarantined to "Policy" (content
filter:russian_text_2)
Thu Sep 10 15:26:08 2015 Info: Message finished MID 164995 done
```

Write a content filter using "Message Language" condition

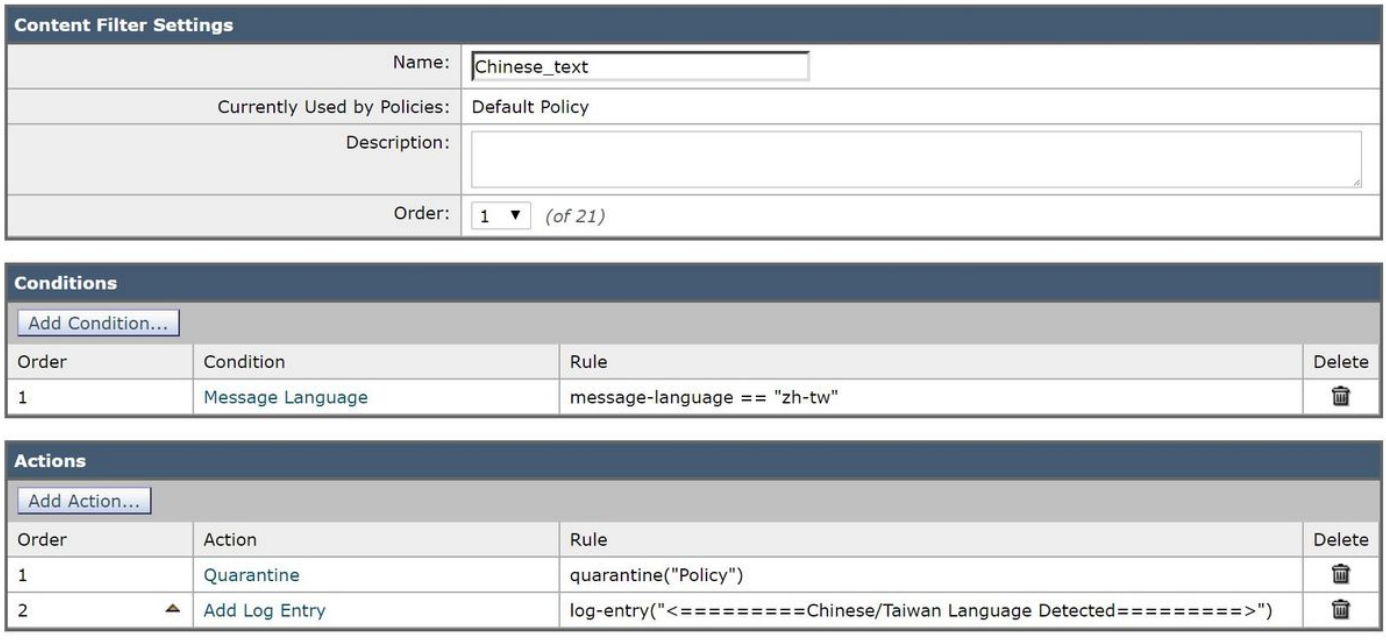
The third option is to use the "message language" condition. The ESA uses the built-in language detection engine to detect the language in a message. The appliance extracts the subject and the message body and passes it to the language detection engine.

The language detection engine determines the probability of each language in the extracted text and passes it back to the appliance. The appliance considers the language with the highest probability as the language of the message. The appliance considers the language of the message as "undetermined" in one of the following scenarios:

- If the detected language is not supported by ESA
- If the appliance is unable to detect the language of the message
- If the total size of the extracted text sent to the language detection engine is less than 50 bytes.

Note: This option is a new feature for AsyncOS Email Security 10.0.0-203 and newer.

The following example will take into account a mail message that contains Chinese/Taiwan based character set. Written as a content filter:



With the content filter configured as above, the mail logs would record similar to the following:

```
Tue Feb 28 06:53:18 2017 Info: Start MID 481 ICID 27
Tue Feb 28 06:53:18 2017 Info: MID 481 ICID 27 From: <end_user@test.com>
Tue Feb 28 06:53:18 2017 Info: MID 481 ICID 27 RID 0 To: <recipient@my_co.com>
Tue Feb 28 06:53:18 2017 Info: MID 481 Subject 'Chinese text test'
Tue Feb 28 06:53:18 2017 Info: MID 481 ready 1047 bytes from <end_user@test.com>
Tue Feb 28 06:53:18 2017 Info: MID 481 matched all recipients for per-recipient policy DEFAULT
in the inbound table
Tue Feb 28 06:53:18 2017 Info: MID 481 interim verdict using engine: CASE spam negative
Tue Feb 28 06:53:18 2017 Info: MID 481 using engine: CASE spam negative
Tue Feb 28 06:53:18 2017 Info: MID 481 interim AV verdict using Sophos CLEAN
Tue Feb 28 06:53:18 2017 Info: MID 481 antivirus negative
Tue Feb 28 06:53:18 2017 Info: MID 481 using engine: GRAYMAIL negative
Tue Feb 28 06:53:18 2017 Info: MID 481 Message language: 'Chinese/Taiwan'
Tue Feb 28 06:53:18 2017 Info: MID 481 Custom Log Entry: <=========Chinese/Taiwan Language
Detected=========>
Tue Feb 28 06:53:18 2017 Info: MID 481 Outbreak Filters: verdict negative
Tue Feb 28 06:53:18 2017 Info: MID 481 quarantined to "Policy" (content filter:Chinese_text)
Tue Feb 28 06:53:18 2017 Info: Message finished MID 481 done
```

References

- Microsoft provides character set names (*.NET name*) in their Code Page Identifiers that can be referenced when writing and configuring filters.

  Note: ANSI code pages can be different on different computers, or can be changed for a single computer, leading to data corruption. For the most consistent results, applications should use Unicode, such as UTF-8 or UTF-16, instead of a specific code page.

- Mozillazine provides in-depth details for Content-type: header, foreign letters, foreign words, and more, in their article for Foreign language spam

Related Information

- **Homoglyph Advanced Phishing Attacks**
- **Cisco Email Security Appliance End User Guides**
- **Technical Support & Documentation - Cisco Systems**