A close up of a logo

Description automatically generated

Webinar FAQ: Tips and tricks for maximizing accuracy and reducing false positive detections in MIP and DLP

May 4, 2022

* **View** all relevant content and resources relating to the compliance solutions in Microsoft 365: <https://aka.ms/mipc/oss>

<https://docs.microsoft.com/en-us/microsoft-365/compliance/communication-compliance-solution-overview?view=o365-worldwide>

* **Sign up** for the MIP Preview Program: <https://aka.ms/MIP-Preview>
* **Follow** us on twitter: [twitter.com/MIPnews](https://twitter.com/MIPnews)
* **Watch** previous webinars: <http://aka.ms/MIPC/webinars>

Q. Where does Syntex fit in here - since it sounds like TC on steroids?

*A. Syntex is a capability within SPO. There is an integration with MIP labels -* [*https://docs.microsoft.com/en-us/microsoft-365/contentunderstanding/apply-a-sensitivity-label-to-a-model*](https://docs.microsoft.com/en-us/microsoft-365/contentunderstanding/apply-a-sensitivity-label-to-a-model)

Q. Where is data processed when submitting docs to trainable classifiers? Can it be specified to ensure it doesn't leave a specific geo for compliance reasons?

*A. For the data storage locations, please see here:* [*https://docs.microsoft.com/en-us/microsoft-365/enterprise/o365-data-locations?ms.officeurl=datamaps&view=o365-worldwide*](https://docs.microsoft.com/en-us/microsoft-365/enterprise/o365-data-locations?ms.officeurl=datamaps&view=o365-worldwide)

Q. When will track and revoke features with combination of unified labeling and SharePoint/OneDrive docs be available? According to the documentation, actually this feature is not supported but it doesn't specify when it will be available

*A. Revoke is in our backlog. We're focusing on the AIP parity items with MIP natively built into Office. There are no timelines to share yet.*

Q. We currently import our EDM's via PowerShell scripts, heard rumor that the process may eventually migrate to allow for uploading of EDM's through the WebUI. Do we know if this will be a future feature and if so, is a timeframe known?

Q. MIP native support of labeling/protecting non-Office files (e.g., PDF, Text files)?

*A. We have a number of PDF related features either in development or preview. We're looking into broader file type support but no commitment at this time.*

Q. Can you export/import rule packages from different tenants?

Q. We have some internal Power BI application, is it possible to get list of files for certain labels quickly (seems search can only return office files)?

*A. Have you looked at Content Explorer inside the Microsoft Purview compliance center? This might give some very quick ways to drill down throughout the organization for labeled content?* [*https://docs.microsoft.com/en-us/microsoft-365/compliance/data-classification-content-explorer?view=o365-worldwide*](https://docs.microsoft.com/en-us/microsoft-365/compliance/data-classification-content-explorer?view=o365-worldwide)

Q. We need some custom business logic based on the information, so we may not use the Explorer directly. Is there a plan to open up API for integration?

A. *We have some guidance here on how to get data from the O365 Management APIs -* [*Microsoft 365 Compliance audit log activities via O365 Management API - Part 1 - Microsoft Tech Community*](https://techcommunity.microsoft.com/t5/security-compliance-and-identity/microsoft-365-compliance-audit-log-activities-via-o365/ba-p/2957171)

Q. Would love to have an AI measure that tells me human resources needs based on the complexity or depth of various DLP integrations

Q. Is there an option to say only flag those with 100% confidence?

*A. There is not really a percent here. You can see the coincidence levels here and you could build your own.* [*Sensitive information type entity definitions - Microsoft Purview | Microsoft Docs*](https://docs.microsoft.com/en-us/microsoft-365/compliance/sensitive-information-type-entity-definitions?view=o365-worldwide)*.*

*To get true 100% match that would be more like an exact data match where you provide the pattern match and we look for that exact match.*

*You can use EDM for this: Get started with exact data match based sensitive information types - Microsoft Purview | Microsoft Docs*

Q. I want to create a rule that detects when an item has 25+ 16 digit numbers that pass the lun test for cc, but not have require keywords. this is to catch the "lists" of SITs being sent out where the supporting elements are not within the 300 character limit

*A. The main thing to be clear on here is how we calculate "Confidence Level" and how you use these. "High Confidence" levels is the highest level we offer in Sensitive Info Types meant for when we have the highest certainty (and in generating a SIT that could be done as well in your definition of it)*

*Again, Exact Data Match is one of the very most certain advanced techniques.*

Q. For data privacy - do we have option to split search base on geo location (Not the Public IP Infor from Azure AD) but through either GPS or proper Geo Tags - Also To query it and combine it with additional data set - either using API - or a query explorer to be specific for a specific requirement -…

*A. We are working on being able to put scoping around what events/policies an an admin can see. We expect to see a preview of this in H2 CY22*

Q. Does SIT creation in PowerShell still use percentages for confidence levels or has that been updated?

Q. Would you recommend keeping the more aggressive rules at higher priority or lower priority?

Q. When we create a DLP policy e.g., deduct certain information and display the notification, is it possible to just show the notification, apparently it also sends an email?

Q. How to make the comparison in the recommendation to keep more aggressive rule with a less aggressive one?

Q. Global Feedback Loop?

Q. I would like to have the possibility to disable those built in SITs that isnaren’tlative (or giving a lot of false/positives) in my tenant/organization

*A. Disabling isn't the best approach. What if you change your mind due to new business requirements later? What is a better approach is filtering out the noise that you don't care about. This is what we're looking at but we don't have a timeline currently.*

Q. How do we evaluate the performance of a new custom/cloned SIT? How do we calculate the accuracy of them without having them in the automatic labeling configuration? Any advice?

*A. You could use simulation mode for auto labeling or use content explorer.*

Q. And for retention labeling and/or retention policies? There is no "simulation" option to check before apply

*A. What conditions are you using to apply your retention labels?*

Q. You solved my question, SITs... then simulation would be the same as with the Labels, isnt it?

*A. Yes this could be used also.*

Q. But is it possible to run a simulation, while the policy is running/enabled?

*A. You could create a new policy for simulation and not enforce it.*

Q. But won’t be easy to add this "simulation" option to the Governance features? Less effort than going back to MIP, re-create label policy...test it..

*A. We are constantly looking at ways to understand the impact of using the classifications of what we define, and certainly this changes slightly depending on the solution being used (example using Retention vs. using Sensitivity labeling...).*

*Understanding how to classification match is made is key, which is the focus of this webinar.*

*We can certainly share the feedback for later improvements and where we can concentrate on.*

Q. For SIT regexes, how can we easily deal with the variety of surrounding characters (space, comma, periods, new line, etc.)?

*A. We publish some of the limitations in the MS Regex engine to help -* [*https://aka.ms/MIPC/RegexValidation*](https://aka.ms/MIPC/RegexValidation)

Q. Is there a way to specifically exclude an SIT if there is a certain keyword detected? For example, a valid credit card number with the phrase "Serial Number" within 100 characters.

Q. Does the MS regex support negative look-ahead?

Q. How can we use “phone number” as a primary element within an EDM SIT?

*A. We don't currently offer an OOB SIT that detects phone numbers, so you would need to create a custom SIT (or a dictionary would work for a small known set) and then map your indexed field to that custom SIT.*

Q. Do you recommend the use of document templates with specific data (or metadata like document properties) and have people use those to base their documents off of - then have SITs that are looking for these pieces of data? To give things a helping hand.

Q. Can we use Anti-Keywords in the condition - "Not Any of These" to minimize the FP's? Will it be effective?

Q. Can you specify unique matching for regex, e.g., email addresses but only unique ones

Q. Why not use a SIT with two different elements for this use case instead? Combine regex in one element and another one with dictionary including the words.