Data & More| For Mircosoft Purview

How to enable Classification, Verification, and Minimizaion In MS Purview with minimum effort!

01 Executive Summary

This white paper is intended for Microsoft professionals who are using (or considering using) Purview to comply with privacy regulations such as GDPR, PIPEDA, CCPA, and HIPAA or to use Purview for DLP or other of its excellent security features.

The add-one Data & More for Purvew enables Microsoft Purview to perform Data Privacy Classification, End-user verification, and Data Minimization out of the box. It works with more the 28 languages and integrates with Purview's own classification, sensitivity labels retention polices.

The whitepaper addresses data classification challenges and the resulting high incidence of false positives and/or missed documents (false negatives), both of which hinder effective deployment.

The paper provides insights and strategies to improve classification accuracy and streamline the application of sensitivity labels, facilitating better compliance and data governance by utilizing the add-on **Data & More for Purview**[1].

Microsoft Purview is an existential GRC tool for managing Microsoft 365, providing organizations with a broad range of management and administrative capabilities. It is important to note that this paper focuses on a small subset of those capabilities, specifically the Purview tools for compliance, which face the innumerable challenges of providing Data Privacy Classification.

Achieving proper compliance hinges on effective data classification, which is a prerequisite for both data governance and data security. Proper classification ensures that data is accurately identified

[1] https://dataandmore.com/for-purview/

and managed, enabling organizations to meet regulatory requirements, protect sensitive information, and maintain robust data governance practices.

Just for the EU and US, this requires hundreds of thousands of combinations of scan patterns, search tokens and keywords to identify all the data covered by Data Privacy Legislation; and this complexity increases exponentially as organizations operate in more languages and additional geographic locations. Adding a third layer of complexity, data privacy scan results need to be validated by data owners and end-users.

Data & More for Purview integrates seamlessly with Purview's compliance components and the unstructured data repositories in Microsoft 365. It provides the needed classification and end-user validation, ensuring that data is handled as it should be—supplemented with reporting for data privacy auditing.

If you want a demo of Data & More, you can try demo.dataandmore.com, it will grant a free scan of up to five users.

You can also send a mail stating that you would like to test Data & More for Purview, and we will provide you with 20 more users to test and a free compliance workshop.

w: dataandmore.com // m: support@dataandmore.com //

Denmark: *+45 4290 1070 - Flaesketorvet 68, 1711 Copenhagen V, Denmark*
Germany: *+49 151 59422362 - Am Steinebrück 29, 40589 Düsseldorf, Germany*
Canada: *+1.587.966.9070 - 500 - 4th Avenue SW, Calgary, Alberta, Canada*

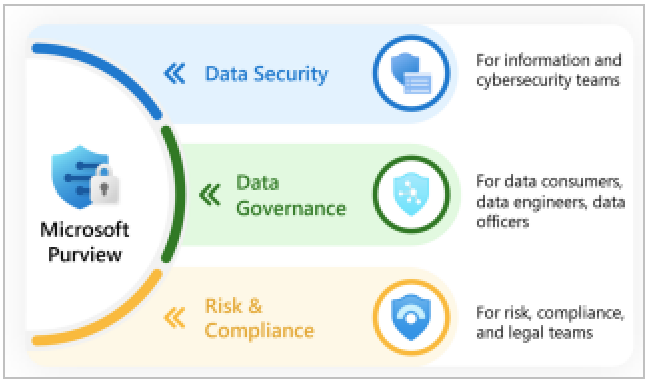
—

Table of Contents

02 Purview Fundamentals[2]

(this part can be skipped if you are familiar with the MS Purview suite)

Purview has become the de facto Governance, Risk, and Compliance (GRC) tool for Microsoft 365, offering a suite of interdependent functionalities. For Purview to operate effectively, it is essential that data within Microsoft 365 is appropriately classified and that sensitivity labels are applied to it. Sensitivity labels ensure robust data governance, enhance data protection, and support compliance with regulatory requirements.


Microsoft Purview
Data Security
For information and cybersecurity teams
Data Governance
For data consumers, data engineers, data officers
Risk & Compliance
For risk, compliance, and legal teams

Data Security

Microsoft Purview provides a robust and coordinated set of data security solutions to help discover and protect sensitive information. The solutions include:

- Data Loss Prevention
- Information Barriers
- Information Protection
- Insider Risk Management
- Privileged Access Management

Data Governance

Microsoft Purview includes unified data governance solutions that help manage data services across on-premises, multi-cloud, and software-as-a-service (SaaS) estate. The solutions include:

- Data Catalog
- Data Estate Insights
- Data Map
- Data Policy
- Data Sharing

[2] Reference: https://learn.microsoft.com/en-us/purview/purview

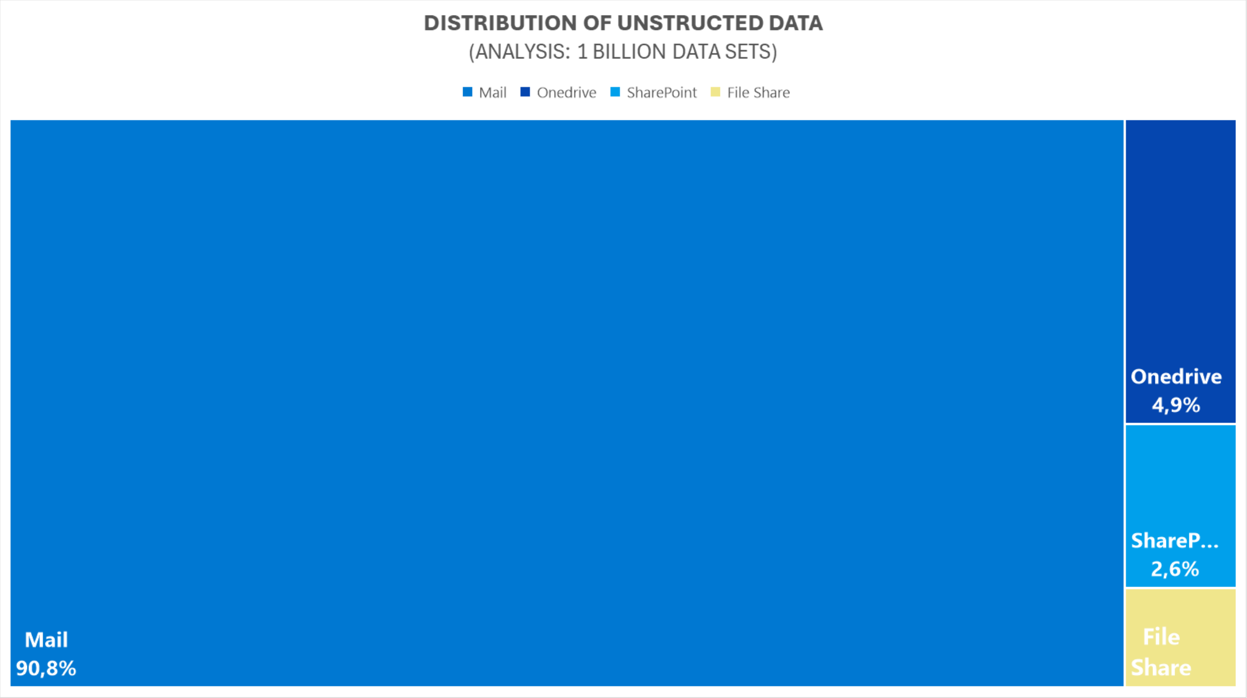
Risk and Compliance

Microsoft Purview includes risk and compliance solutions to help your organization minimize compliance risks and meet regulatory requirements. The solutions include:

- Audit
- Communication Compliance
- Compliance Manager
- Data Lifecycle Management
- eDiscovery

03 Purview – What can be protected and classified by Purview (and what can not)

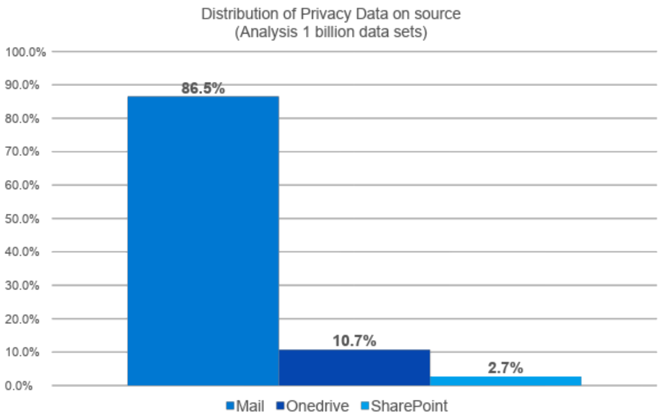
Data & More analyses billions of datasets for our customers every day. To assess how much data Purview can classify given the inherent restrictions in Purview, we analyzed a sample of 1 billion datasets from different industries, which all used Microsoft 365 as their primary communication platform.

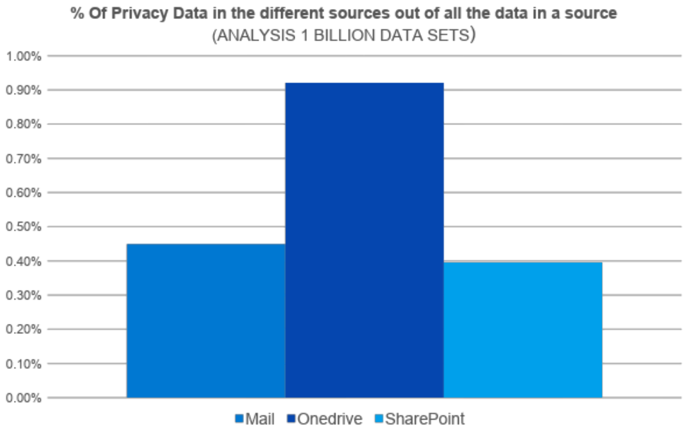


Where is privacy data located in Microsoft 365

If we start by not including file shares (contains 1,37% of privacy data) - since it can’t be protected directly by purview and thus must be handled by another solution - we have Mail,   and Sharepoint. (Teams documents are stored in Sharepoint and OneDrive)

Unsurprisingly, the vast majority (+86%) of privacy data is in emails or attachments to emails. If we analyze the dataflow, more than 82% of privacy data are first identified in email and then later added to Onedrive or Sharepoint.



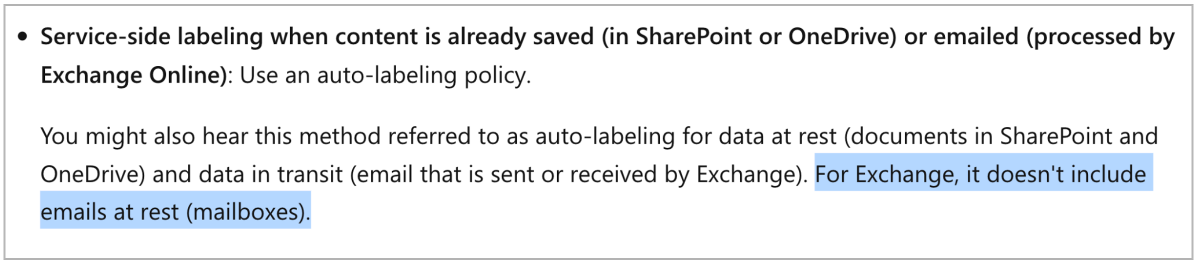


If we look at privacy density per data source, OneDrive is where privacy documents occur most frequently.

What data can be classified and what can be protected by Purview

Two primary factors determine whether a file can be classified and/or protected: the document's location and file type.

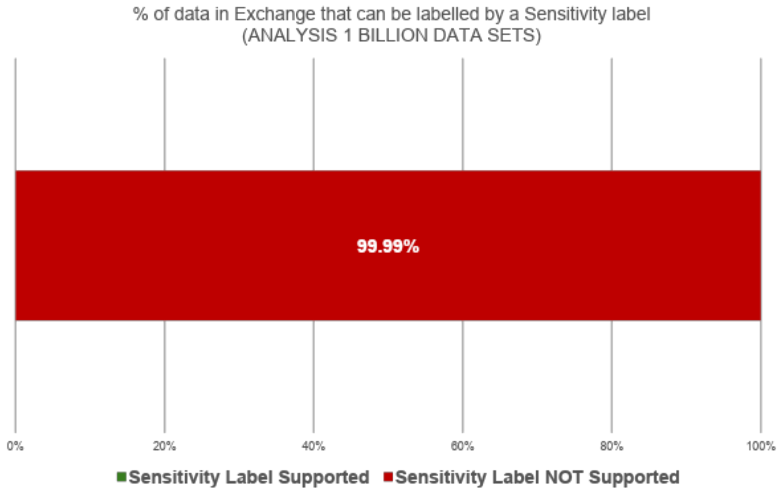
Locations are the most critical factor, as they exclude all the data stored in Exchange. When it comes to Purview, it can only classify data "in transit and not at rest."



- **Service-side labeling when content is already saved (in SharePoint or OneDrive) or emailed (processed by Exchange Online)**: Use an auto-labeling policy.

  You might also hear this method referred to as auto-labeling for data at rest (documents in SharePoint and OneDrive) and data in transit (email that is sent or received by Exchange). For Exchange, it doesn't include emails at rest (mailboxes).

https://learn.microsoft.com/en-us/purview/apply-sensitivity-label-automatically?view=o365-worldwide[3]



From a data protection and privacy perspective, this is a significant red flag for most privacy projects. As we recall, the majority of privacy data is located in Mail. Purview cannot classify, protect, or clean your old Mails.

Now, what you can do—if you get your classification up and running (and get it perfect from the beginning), then you can set auto labeling to catch data in and out of the Exchange server.

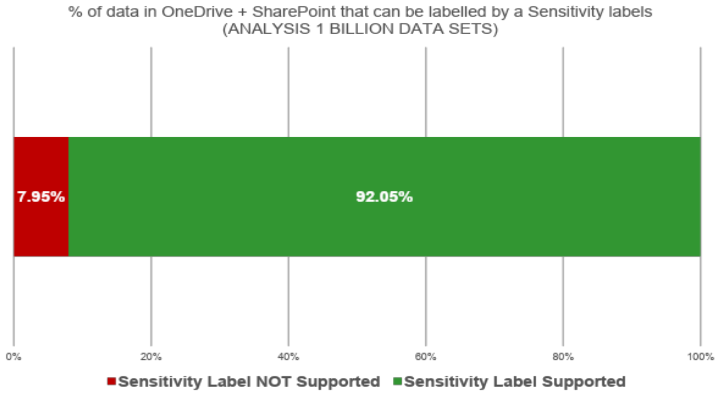
With time (and if you delete all old mail), you could get a compliant Exchange. However, this is a theoretical approach and is not recommended for practical application.

A more realistic approach is to use a privacy compliance tool that works with Exchange out of the box, such as Data & More for Exchange Online.

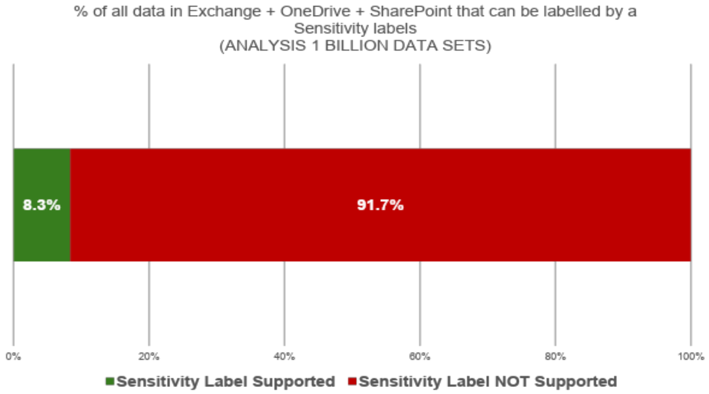
Some file types can be classified but not protected; others can be classified and protected; and again, some can only be protected. For a complete list of which files are eligible for classification and/or protection, please refer to the Microsoft documentation.

[3] https://learn.microsoft.com/en-us/purview/apply-sensitivity-label-automatically?view=o365-worldwide

When it comes to OneDrive and SharePoint, Purview, on the other hand, excels and can classify 92% of the privacy data— if the classification has been built correctly.

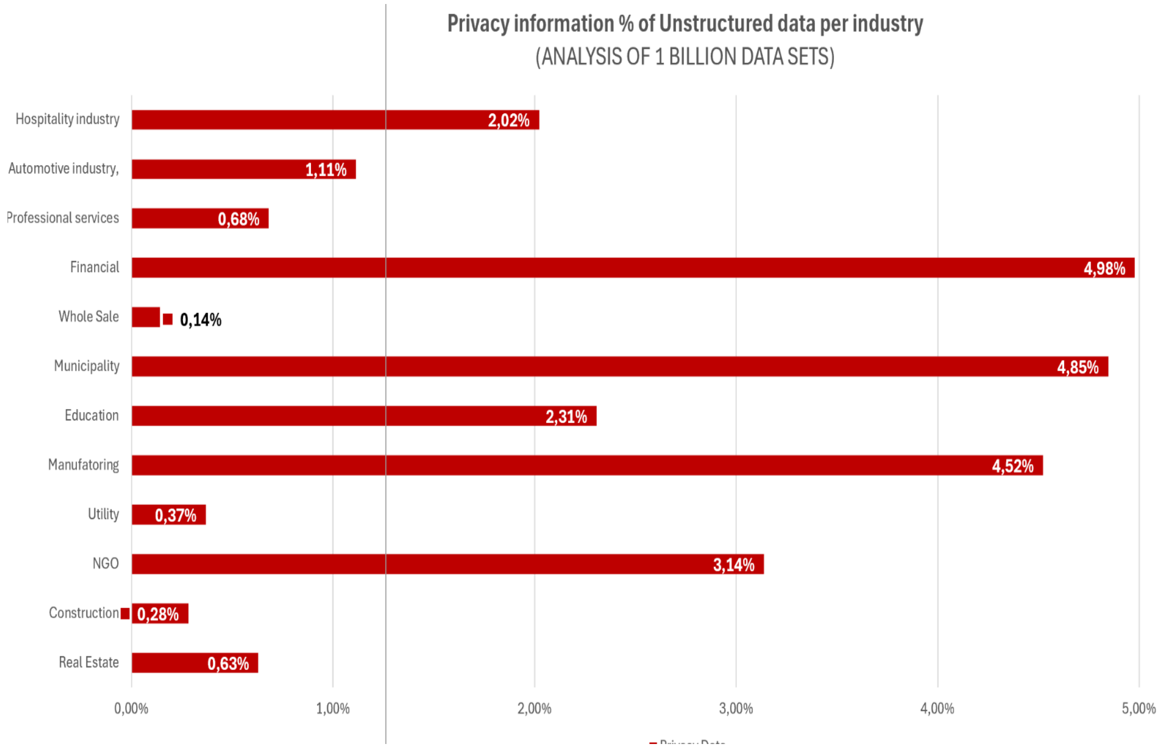


Let's combine the locations and file types for Privacy Data. In that case, we find that **only around 8% of the Privacy Data in Microsoft 365 would be protected if we implemented Purvew today**—given that the Purview classification was perfect.



Privacy data are not created equal - among industries

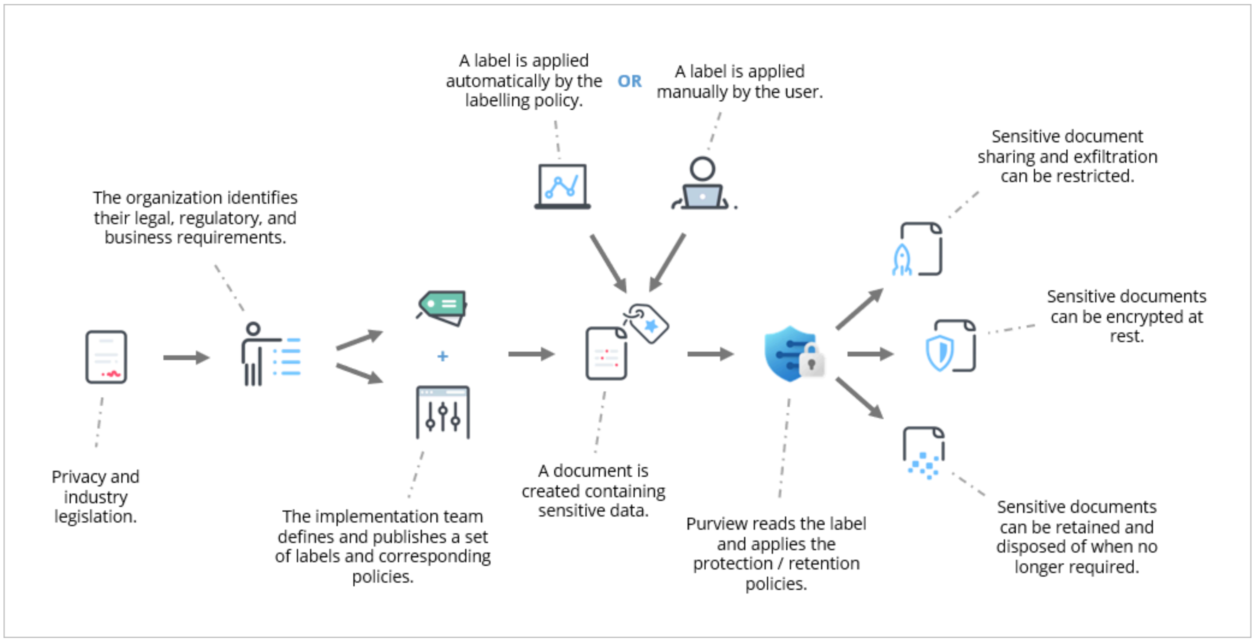
All industries are creating privacy data - some are creating data as a part of the client and customer interactions, such s B2C, where B2B often has privacy data related to employees and recruitment.



04 Sensitivity Labels – The Rosetta Stone of Purview

The way Purview tracks data in Microsoft 365 is by applying a small metadata code to each document. The metadata's names have changed over time but are usually called Sensitivity Labels or Microsoft Information Protection labels - MIP Labels in short. The terms “Sensitivity Labels” and Microsoft Information Protection Labels (MIP) are often used interchangeably. You can find the label overview in your Microsoft 365 Compliance Center under Purview - Information Protection - Labels.

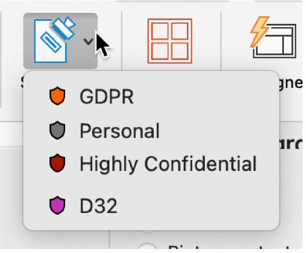
The general idea is that Purview can be configured to protect data and comply with legislation based on a specific rule set, such as GDPR, HIPAA, or PIPEDA, following the model below*.


A label is applied automatically by the labelling policy.
OR
A label is applied manually by the user.
The organization identifies their legal, regulatory, and business requirements.
Sensitive document sharing and exfiltration can be restricted.
Sensitive documents can be encrypted at rest.
Privacy and industry legislation.
The implementation team defines and publishes a set of labels and corresponding policies.
A document is created containing sensitive data.
Purview reads the label and applies the protection / retention policies.
Sensitive documents can be retained and disposed of when no longer required.

*The critical point in this flow is how to map a given piece of legislation to a label and how the label is applied to the document. There are two ways this happens.

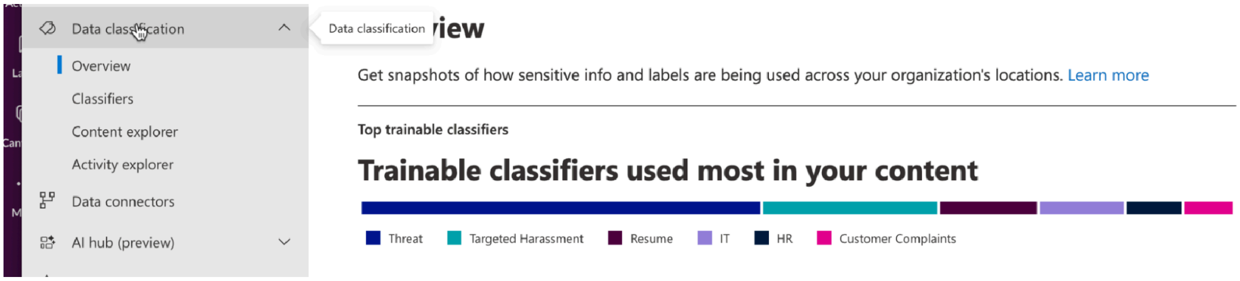
Manual Classification

Purview offers a smooth in-app classification experience for the end user. Sensitivity labels can be applied via the office app when the document is being created or saved. Manual classification has been around for several years and is the most reliable way to classify data in Purview - if the employee chooses the right label.



Automatic Classification

In addition to manual classification, Purview offers the option to use automatic classification, where the user can create document classes by using several RegExs, libraries, and some trainable classifiers provided by Microsoft. However, as we shall see, this isn't sufficient to meet Data Privacy legislation requirements and requires a dedicated team (and some more advanced tools) to be successful.



Get snapshots of how sensitive info and labels are being used across your organization's locations. Learn more

Top trainable classifiers

Trainable classifiers used most in your content

05 Privacy Classification with Purview

Privacy classification is a subset of data classification and is concerned with identifying sensitive personal information. What constitutes person-related and sensitive information can vary from jurisdiction to jurisdiction.

Generally speaking, privacy legislation is based on the idea that a person (or 'data subject') has the right to their data, and organizations have a responsibility to delete personal data when they no longer have a legitimate purpose to keep it.

From a legal perspective, the object of interest is the Data Subject, but this "Subject" must be created based on data analyzed either at the document level (document classification) or within the specific text (token classification).

For Privacy Classification to be effective, all three levels of analyses must be applied. The Data Subject Model provides the legal reference needed by Data Privacy Laws. The document classification is used to identify documents that must be moved, encrypted, or deleted. The token-based analyses are used to construct both the document classification and the identification of the Data Subject.

The Data Subject Model for Privacy Classification

The key principle of the Data Subject Model (DSM) is to identify each data subject and link all personal data for that 'real person' to the data subject. It must also have a way of verifying that the identified data is, in fact, personal data that belongs to the data subject. This makes data discovery, verification, and remediation much easier, especially when it needs to be applied to all data for a specific person.

Another significant advantage of using a Data Subject Model for privacy classification is that it allows the use of multiple vital identifiers to create a comprehensive privacy profile for each data subject. This approach ensures that all relevant personal data is accurately associated with the correct individual, enabling more effective management of privacy requests and compliance with data privacy regulations.

The key principle of the Data Subject Model (DSM) is to identify each data subject and link all personal data for that 'real person' to the data subject. It must also have a way of verifying that the identified data is, in fact, personal data that belongs to the data subject. This makes data discovery, verification, and remediation much easier, especially when it needs to be applied to all data for a specific person.

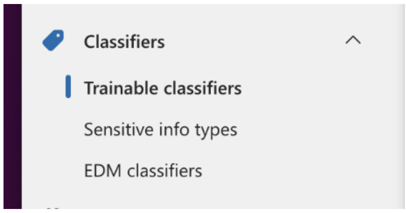
Another significant advantage of using a Data Subject Model for privacy classification is that it allows the use of multiple vital identifiers to aggregate a comprehensive privacy profile. This approach ensures that all relevant personal data is accurately associated with the correct individual, enabling more effective management of privacy requests and compliance with data privacy regulations.

By leveraging this model, organizations can enhance their ability to manage data subject rights and ensure that personal data is properly classified, protected, and governed throughout its lifecycle.

Microsoft Purview does not offer a Data Subject Model, so Purview Developers must either create and maintain their own DSM or use the DSM provided by Data & More for Purview (see below).

Purview’s Analytics Focus

Purview's analytical focus is either on the document or text/token level. Purview offers three different classification tools: “Trainable Classifiers'' for document classification, “Sensitive Information types,” and EDM (Exact Data Match) Classifiers for token classification.



Trainable Classifiers (Document Classification)

Microsoft Purview trainable classifiers are pre-trained to recognize specific types of sensitive content. As of June 2024, there are 136 trainable classifiers available in Microsoft Purview. From a Data Privacy Perspective, the following are relevant—but unfortunately far from enough.

HR (GDPR and Privacy Regulations)	Healthcare (HIPAA)
Employee Disciplinary Action Files Employee Insurance Files Employment Agreements Paystubs	Health and Medical Forms

For more detailed information on trainable classifiers and their setup, you can visit the Microsoft Learn website and explore the relevant sections on data classification and information protection in Microsoft Purview (Microsoft Learn[4]).

So, to use Purview for privacy classification, you require a complete set of classification rules that can find all the privacy data you are obligated to manage. This set of rules either needs to be built or licensed from someone who has already created them.

Privacy Classification That Must be Built, Bought, or Borrowed

To be compliant, the identification of Privacy Data has to be extensive to be able to match the requirements from the local data authorities. The solution must identify and map real-world privacy

[4] Reference: https://learn.microsoft.com/en-us/purview/trainable-classifiers-get-started-with

data in the data sources. See the list of the required privacy classifications to be compliant with privacy legislation: https://support.dataandmore.com/en/knowledge/what-is-gdpr-data-classification

Purview only has examples of privacy classification, and they only cover parts of the legally required privacy classification. Below are examples of types of privacy classification that are not included in Purview. For Purview to be able to find the legally required types of privacy data, all the Purview classification examples need to be tested, validated, and configured to work on the customers' data, otherwise, it is not precise enough and results in false positives and negatives - and missed non-compliant data. In addition, the customer must make a gap analysis and then create all the missing classifications to have a compliant privacy classification.

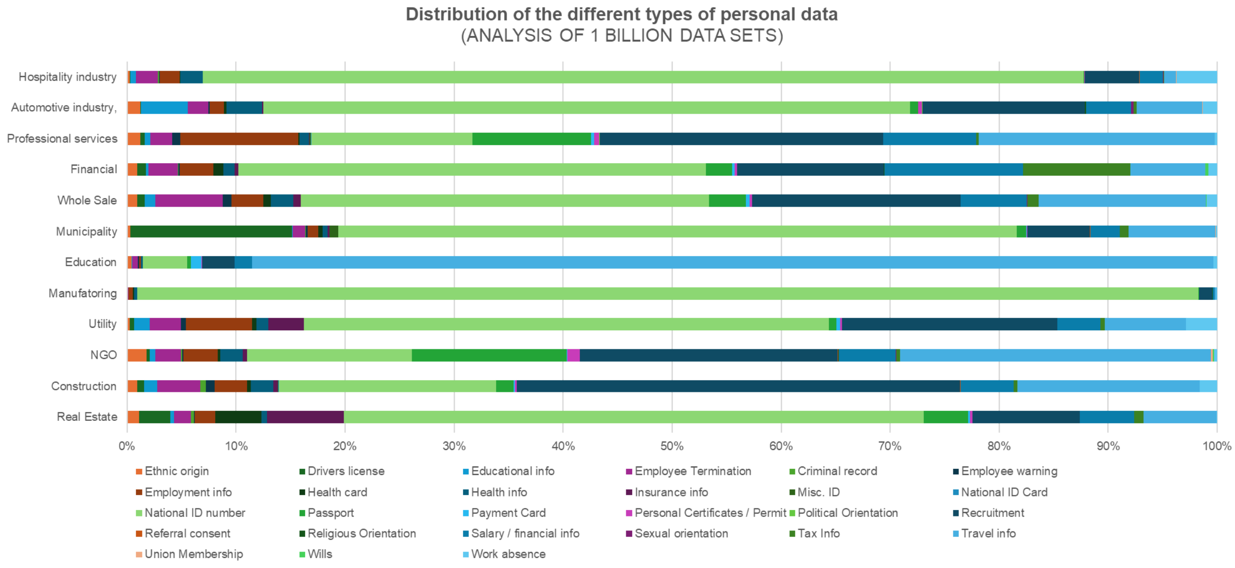
Examples of types of required privacy classification that are not included in Purview

- Political orientation, Sexual orientation, Ethnic origin, Trade union affiliation, Religious orientation
- All types of health information, medicine, diagnosis
- Personal tax information
- Salary information
- Employment information
- Recruitment information
- Written warnings
- Work absence
- Criminal records
- Written consents
- Travel information
- Photo geolocation

And so on... (Data & More have identified over ten thousand different searches / document types)

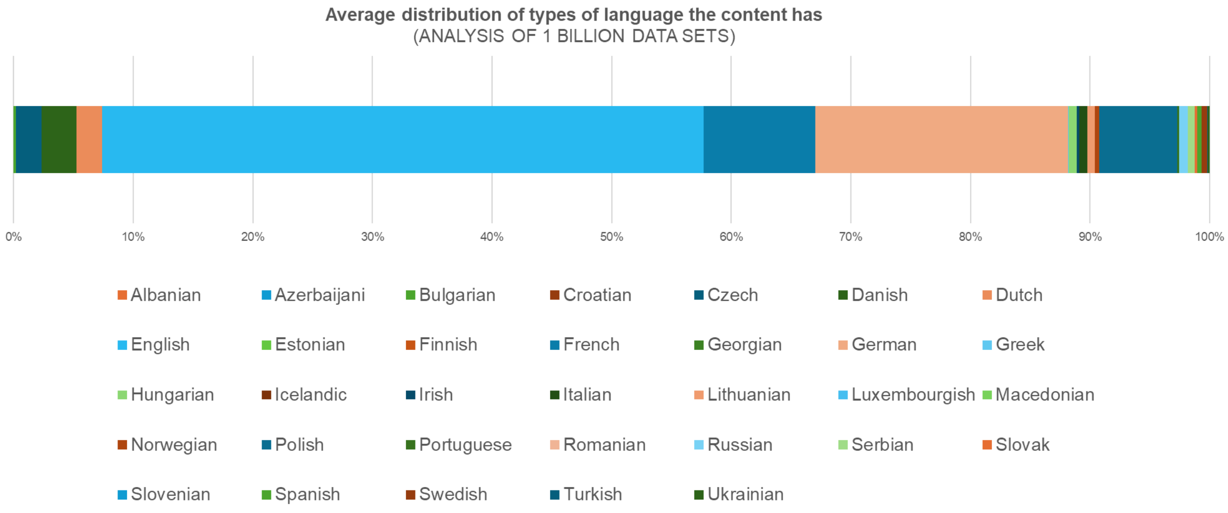
One of the complications with extensive Privacy classification is that the Privacy Data requirements are specific to each location, country, jurisdiction and industry.. For example, hospital locations vary, union names differ, and religious institutes have different names in different countries.

Each industry has different types of personal data

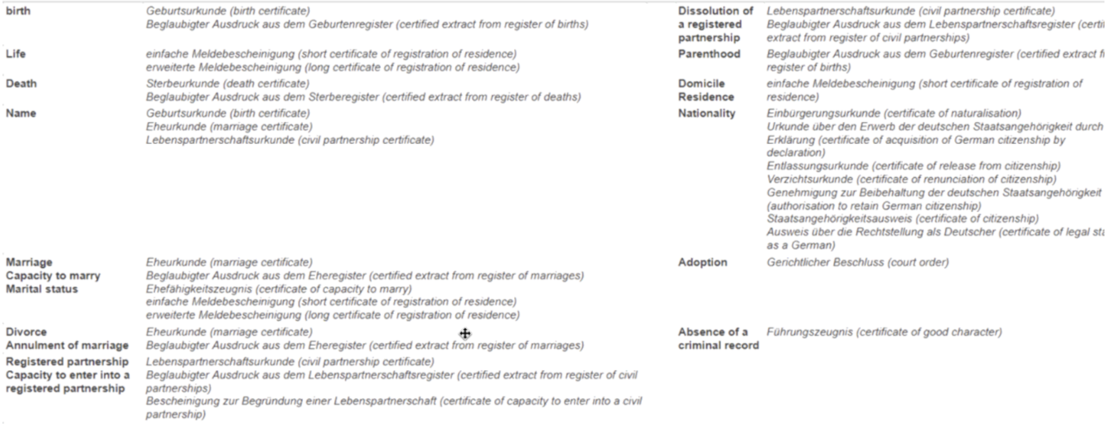


International organisations also cover different countries and thereby content will be in a lot of different languages. Below is an overview of the distribution of various languages. This is important because any classification has to be able to identify content in the different languages.

Distribution of types of languages in the content



In addition to the complexity of the examples above, some countries have decided to make their own list of what documents constitute Privacy Data, such as Germany's example below.



birth	*Geburtsurkunde (birth certificate)* *Beglaubigter Ausdruck aus dem Geburtenregister (certified extract from register of births)*
Life	*einfache Meldebescheinigung (short certificate of registration of residence)* *erweiterte Meldebescheinigung (long certificate of registration of residence)*
Death	*Sterbeurkunde (death certificate)* *Beglaubigter Ausdruck aus dem Sterberegister (certified extract from register of deaths)*
Name	*Geburtsurkunde (birth certificate)* *Eheurkunde (marriage certificate)* *Lebenspartnerschaftsurkunde (civil partnership certificate)*
Marriage **Capacity to marry** **Marital status**	*Eheurkunde (marriage certificate)* *Beglaubigter Ausdruck aus dem Eheregister (certified extract from register of marriages)* *Ehefähigkeitszeugnis (certificate of capacity to marry)* *einfache Meldebescheinigung (short certificate of registration of residence)* *erweiterte Meldebescheinigung (long certificate of registration of residence)*
Divorce **Annulment of marriage**	*Eheurkunde (marriage certificate)* *Beglaubigter Ausdruck aus dem Eheregister (certified extract from register of marriages)*
Registered partnership **Capacity to enter into a registered partnership**	*Lebenspartnerschaftsurkunde (civil partnership certificate)* *Beglaubigter Ausdruck aus dem Lebenspartnerschaftsregister (certified extract from register of civil partnerships)* *Bescheinigung zur Begründung einer Lebenspartnerschaft (certificate of capacity to enter into a civil partnership)*
Dissolution of a registered partnership	*Lebenspartnerschaftsurkunde (civil partnership certificate)* *Beglaubigter Ausdruck aus dem Lebenspartnerschaftsregister (certi extract from register of civil partnerships)*
Parenthood	*Beglaubigter Ausdruck aus dem Geburtenregister (certified extract fi register of births)*
Domicile **Residence**	*einfache Meldebescheinigung (short certificate of registration of residence)*
Nationality	*Einbürgerungsurkunde (certificate of naturalisation)* *Urkunde über den Erwerb der deutschen Staatsangehörigkeit durch Erklärung (certificate of acquisition of German citizenship by declaration)* *Entlassungsurkunde (certificate of release from citizenship)* *Verzichtsurkunde (certificate of renunciation of citizenship)* *Genehmigung zur Beibehaltung der deutschen Staatsangehörigkeit (authorisation to retain German citizenship)* *Staatsangehörigkeitsausweis (certificate of citizenship)* *Ausweis über die Rechtstellung als Deutscher (certificate of legal sta as a German)*
Adoption	*Gerichtlicher Beschluss (court order)*
Absence of a criminal record	*Führungszeugnis (certificate of good character)*

So, if you are an international organization, you must build and maintain more than 10,000 Trainable queries / Classes in Purview, or around 800+ for a single local jurisdiction in multiple languages. These must all be tested, adjusted, quality-assured, and approved before they can be effectively used on real data.

Given that Purview does not support dynamic reclassification of data without a rescanning of data, any validation and testing of classification of just a single query on a big enough dataset to be able to validate the accuracy of the classification will take weeks (or longer). The effort to create thousands of queries will take years of testing and validating, also because some queries are part of other queries and therefore the quality has to be high enough for the individual components for the classification to work.

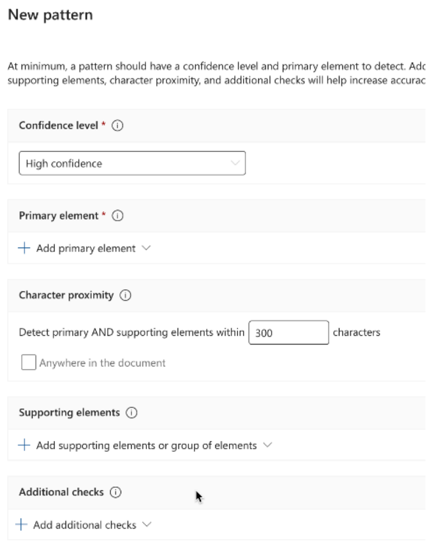
In addition, the creation and validation have to be conducted in different languages, and on different data sets and therefore requires a qualified multilanguage classification team working on this over a long period.

The alternative is to utilize Data & More for Purview, which provides all the privacy classification rules that are required and includes the ongoing maintenance of those rules. It also provides the needed searches and labelling.

Many organizations get stuck with Purview sooner or later regarding classification and clean-up, and reach out to Data & More (either to help succeed with Purview or go with Data & More alone). As Data & More provides comprehensive classification and data handling out-of-the-box, it is often an easy decision for the client to test Data & More's solution on real data, as this requires only a couple of hours of involvement for the customer.

Sensitive Information Types (SIT)

Purview offers SITs for Passports, Driver's Licenses, Social Security Numbers, and Addresses for nearly all countries. As shown in the picture to the right, Purview provides a nice interface for defining SITs.



New pattern

At minimum, a pattern should have a confidence level and primary element to detect. Adc
supporting elements, character proximity, and additional checks will help increase accurac

Confidence level * ⓘ

High confidence

Primary element * ⓘ

+ Add primary element

Character proximity ⓘ

Detect primary AND supporting elements within 300 characters

□ Anywhere in the document

Supporting elements ⓘ

+ Add supporting elements or group of elements

Additional checks ⓘ

+ Add additional checks

The SIT is implemented as regular expressions, which unfortunately results in a very high volume of both false positives and false negatives.

False positives (data that is identified as an SIT but is, in fact, not) are very often linked to long sequential numbers like SKU, Invoice, and VAT numbers. As an example, for any company operating in the European Union, a 10-digit invoice number will, with a 25% likelihood, be identified as a Danish Social Security Number. As invoices are often much more prevalent in the data sources, false positives can become much more likely than an actual SIT.

Many organizations try to address this challenge by implementing increasingly complex regular expressions. However, this approach has a significant impact on processing time and usually provides only minimal gains in accuracy.

For SIT to be useful, the regular expression search must be combined with other information types, such as text language, full name, or EDM, to increase the likelihood of the text containing actual Privacy Data, which requires complex rules and extensive testing.

Exact Data Match (EDM)

Exact Data Match (EDM) in Microsoft Purview is a feature designed for precise data classification by matching exact values from pre-defined datasets. This method is distinct from traditional pattern-based sensitive information types (SITs) and significantly reduces false positives by ensuring matches are based on exact data values. However, relying on an “Exact Data Match” increases the risk of false negatives, aka not identifying a text / token as Privacy Data that should have been identified.

EDM Functionality

1. EDM utilizes exact matches, reducing false positives compared to pattern-based detection. It is effective for structured data, such as customer records, employee information, and financial data.

2. Users define the schema, including primary and secondary data fields that identify sensitive items. This involves creating a CSV file with the sensitive data fields, which are then hashed and uploaded.
3. EDM supports regular updates, allowing the sensitive datasets to be refreshed without extensive reconfiguration, ensuring that the classification remains current
4. EDM integrates with compliance and data loss prevention (DLP) policies within Microsoft Purview, including auto-labeling for sensitivity labels and retention policies.

Implementation Steps

1. **Define the Schema**. Establish a schema for the exact data match, specifying primary and secondary elements that need to be matched. One key challenge when working with EDM is that people tend to spell the exact same word in different ways and with varying notes of support, which leads to false negatives.
2. **Prepare and Upload Data**. Export sensitive data into a CSV file, hash the data and securely upload it using the EDM Upload Agent.
3. **Configure and Test**. Set up the EDM in the Microsoft Purview Compliance Portal, configure detection rules, and test the classifier to ensure accurate identification of sensitive data.
4. **Deploy and Monitor**. You can now use the EDM as a building block in a hierarchy classification and deploy and monitor the EDM propagation in the tenet.

It's also important for any organization going through this implementation process with Purview to realize that each change to any of the classifiers - SIT or EDM - requires a full re-scan of the data. For organizations with large amounts of data, this results in a significant time gap between the time a change is made and the time classification results including that change are available. For this reason, many organizations tend to queue and prioritize changes as they refine their data discovery to reduce the number of full data scans that are required.

The Importance of End-User Validation Before Data Minimization

Correct Privacy Classification is crucial before encrypting, moving, and deleting data. However, if data is misclassified, it can severely disrupt efficient data management. Therefore, IT departments must involve end-users (the data owners) in the validation process before encrypting or deleting data.

Risks of Misclassification

- **Operational Disruption**: Misclassified data can lead to the unintended encryption or deletion of critical information. For example, if a CFO loses spreadsheets needed for an upcoming board meeting or if the sales team loses a draft contract, it can significantly impact business operations.
- **User Frustration**: False positives in data classification may cause significant inconvenience for end-users. If HR cannot find a crucial CV and wrong ones are presented instead (maybe not even a CV), the hiring processes may be delayed. Such issues can lead to many

complaints directed at the IT department, undermining trust and cooperation between IT and other departments.

Role of End-User Validation

- **Accurate Classification**: Data owners are most familiar with their data and can validate the classification accurately. Involving them ensures that only the appropriate data is encrypted or deleted, minimizing the risk of misclassification.
- **Data Minimization**: End-user validation helps precisely identify data that genuinely needs to be kept, ensuring that valuable data is retained while redundant or unnecessary data is appropriately minimized.
- **Data Prioritization**: End-users, in particular data owners and data stewards are in the best position to identify the most important sets of data for the organization. This allows the organization to prioritize management and remediation efforts on the data where it will have the greatest impact.

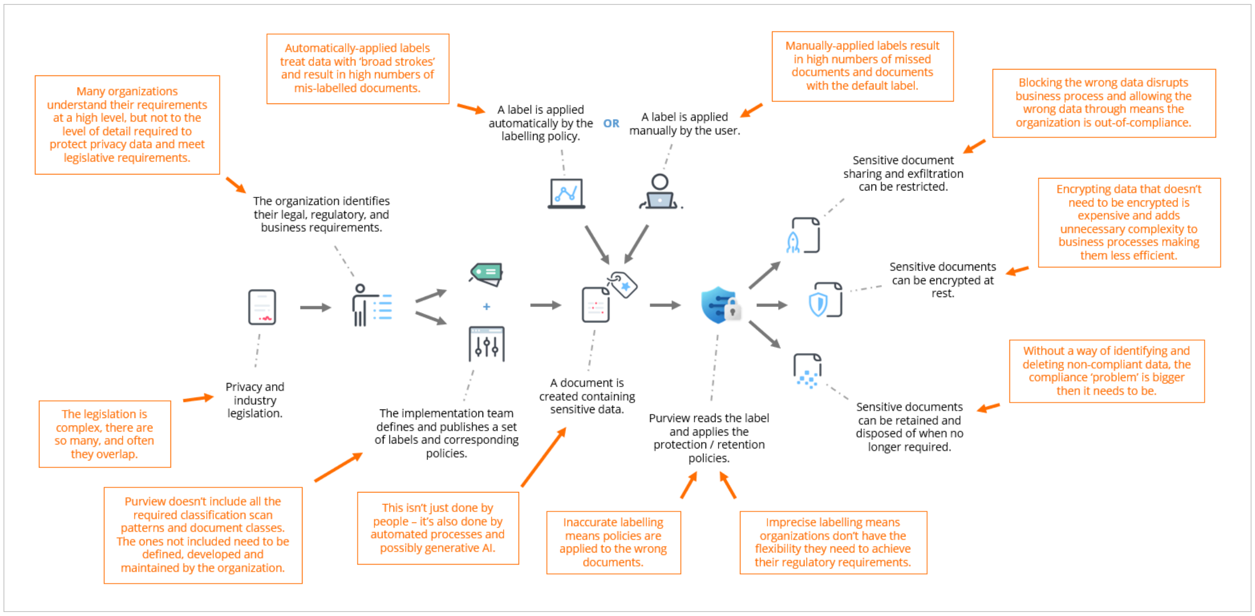
Feasibility

- **Manageable Data Volume**: If the volume of privacy data is very low data owners can manually validate the classification, provided they have the right tools. This process, though requiring some effort, ensures data integrity and operational continuity. End-users can use Purview for validation, which offers tools to assist in the classification process.
- **For Larger Data Volumes (+100 per user)**: If the data volume is large, it is recommended to build an end-user validation overview or use tools like "Data & More for Purview," which can automate, streamline and handle the validation process for larger datasets and numerous end-users.

Involving data owners in the validation process before data minimization is critical. It ensures accurate data classification, prevents operational disruptions, and maintains user trust. The effort required for manual validation is justified by the significant benefits of avoiding misclassification and ensuring efficient data management.

The Purview Challenge

Microsoft Purview is a powerful tool for Governance, Risk, and Compliance (GRC) in Microsoft 365. To ensure compliance with data privacy regulations, extensive privacy classification and a tool for end-user validation must be integrated with Purview. Below, in orange, are the key challenges with the Purview methodology.


Many organizations understand their requirements at a high level, but not to the level of detail required to protect privacy data and meet legislative requirements.
Automatically-applied labels treat data with 'broad strokes' and result in high numbers of mis-labelled documents.
A label is applied automatically by the labelling policy.
OR
A label is applied manually by the user.
Manually-applied labels result in high numbers of missed documents and documents with the default label.
Blocking the wrong data disrupts business process and allowing the wrong data through means the organization is out-of-compliance.
Sensitive document sharing and exfiltration can be restricted.
Encrypting data that doesn't need to be encrypted is expensive and adds unnecessary complexity to business processes making them less efficient.
The organization identifies their legal, regulatory, and business requirements.
Sensitive documents can be encrypted at rest.
+
Without a way of identifying and deleting non-compliant data, the compliance 'problem' is bigger then it needs to be.
Privacy and industry legislation.
A document is created containing sensitive data.
Sensitive documents can be retained and disposed of when no longer required.
The legislation is complex, there are so many, and often they overlap.
The implementation team defines and publishes a set of labels and corresponding policies.
Purview reads the label and applies the protection / retention policies.
Purview doesn't include all the required classification scan patterns and document classes. The ones not included need to be defined, developed and maintained by the organization.
This isn't just done by people – it's also done by automated processes and possibly generative AI.
Inaccurate labelling means policies are applied to the wrong documents.
Imprecise labelling means organizations don't have the flexibility they need to achieve their regulatory requirements.

Privacy Classification Maintenance

Developing and maintaining data privacy classifications is crucial. These classifications must be updated weekly or monthly to account for changes in compliance rulings and various types of information, such as diseases and sexual orientations. Managing and maintaining over 250,000 different entities is necessary to effectively identify non-compliant privacy data. For an organization to comply with privacy legislation, it must accurately identify over 95% of privacy data across different data sources. Failure to do so can cause non-compliance with privacy laws.

End-User Validation

End-user validation is critical for ensuring the accuracy of data classifications. Users familiar with the content can verify classifications, reducing the likelihood of misclassification. For larger datasets (over 100 items per user or more than 50 end-users), it is recommended to build an end-user validation report or use tools like Data & More for Purview to streamline the validation process. This approach ensures accurate data classification and compliance with privacy regulations while minimizing disruptions and employee frustration.

Custodian Mapping for End-User Validation

To perform end-user validation, an "End-User" must be appointed as the data owner. For email and OneDrive, this is the account owner, but for SharePoint and shared mailboxes, this is often not obvious. In SharePoint, a site can have "Owners" and "Members," but the "Owner" of a site can be a system or an IT administration that has no responsibility for the data itself, or an Owner might not longer be in the organization. The same issues apply to shared mailboxes - where no manager is appointed.

End-users appointed to oversee the compliance of a data repository are often referred to as "Custodians," and the process of assigning Custodians to data is referred to as Custodian Mapping. If you are in a smaller organization, you can conduct Custodian mapping with a combination of Excel, PowerShell, and Entra. Still, for larger organizations where the act of Custodian Mapping must be delegated within the organizations, it is recommended to use dedicated Custodian Mapping tools such as the one provided in D&M for Purview[5].

Out-of-the-Box Limitations

Purview provides some classification examples out of the box, but they cover only a small fraction of the data privacy requirements mandated by laws like GDPR and PIPEDA. The default classifications in Purview often produce many false positives and negatives, necessitating significant enhancements and updates by the customer.

Maintaining extensive and accurate privacy classification is a complex and ongoing task. Many organizations abandon auto-classification, relying on end-users to manually add sensitivity labels to documents. However, correct and extensive automatic classification can significantly improve Purview functionalities like Data Loss Prevention (DLP) and Data Minimization, thereby enhancing security and reducing the risk of data breaches, regulatory penalties, and identity theft.

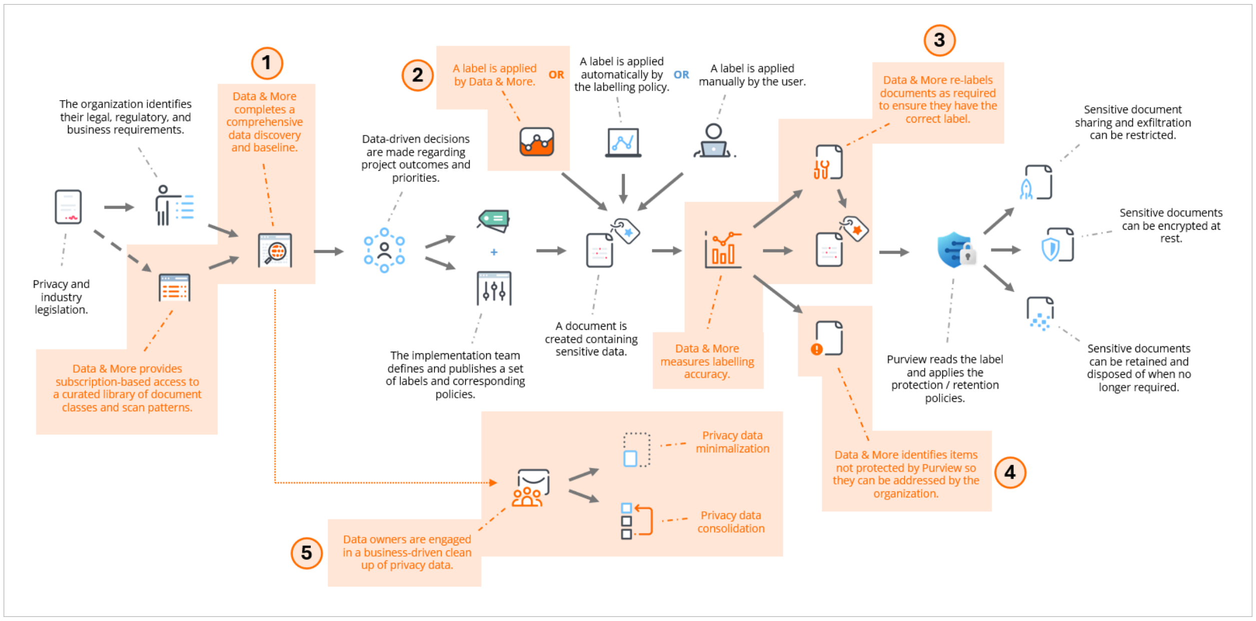
[5] https://dataandmore.com/for-purview/

06 Data & More for Purview Fundamentals

Data & More for Purview (DMCS-PW) is an add-on to MS Purview that enables **Extensive Privacy Classification** for GDPR, HIPAA, CCPA, and PIPEDA, as well as **End-User Validation** and **Custodian Mapping**.

It is used in addition to any auto-classification and manual classification conducted in Purview. DMCS-PW works by adding Sensitivity labels to the content in Microsoft 365. Then, Purview can be used for DLP, Data Mitigation, and all the other GRC functionalities that make Purview great.

From a compliance perspective, the use of DMCS-PW is highlighted in orange in the illustration below:



The organization identifies their legal, regulatory, and business requirements.
Privacy and industry legislation.
1
Data & More completes a comprehensive data discovery and baseline.
Data & More provides subscription-based access to a curated library of document classes and scan patterns.
Data-driven decisions are made regarding project outcomes and priorities.
The implementation team defines and publishes a set of labels and corresponding policies.
2
A label is applied by Data & More.
OR
A label is applied automatically by the labelling policy.
OR
A label is applied manually by the user.
A document is created containing sensitive data.
Data & More measures labelling accuracy.
3
Data & More re-labels documents as required to ensure they have the correct label.
4
Data & More identifies items not protected by Purview so they can be addressed by the organization.
Purview reads the label and applies the protection / retention policies.
Sensitive document sharing and exfiltration can be restricted.
Sensitive documents can be encrypted at rest.
Sensitive documents can be retained and disposed of when no longer required.
5
Data owners are engaged in a business-driven clean up of privacy data.
Privacy data minimalization
Privacy data consolidation

The five key capabilities, corresponding to the numbers on the diagram, are:

1. **Comprehensive data discovery and classification**. The 'secret sauce' is Data & More's curated and maintained library of document classes and scan patterns. They are used to identify and classify *all* privacy data and enable data-based planning and decision-making during the implementation of the various Purview capabilities.
2. **Applying labels based on Data & More classification**. The labels being applied are Microsoft (MIP) labels, enabling all the functionality of Purview with the accuracy and completeness of Data & More data discovery and classification.
3. **Audit and correct documents that aren't labeled properly**. Data & More for Purview identifies and can address any files mis-labeled by Purview auto-labeling policies and user error. Data & More for Purview also facilitates bulk re-labeling should business requirements change over time.

4. **Understand the data NOT protected by MIP**. Data & More for Purview can report on files containing privacy data that are *not* protected by Purview and quantify the level of data compliance risk created by that gap'.
5. **Provide a mechanism for cleaning up privacy data**. Data & More for Purview facilitates the business-driven clean-up of non-compliant data to minimize the privacy data stored and consolidate where privacy data is stored, simplifying both compliance and on-going administration.

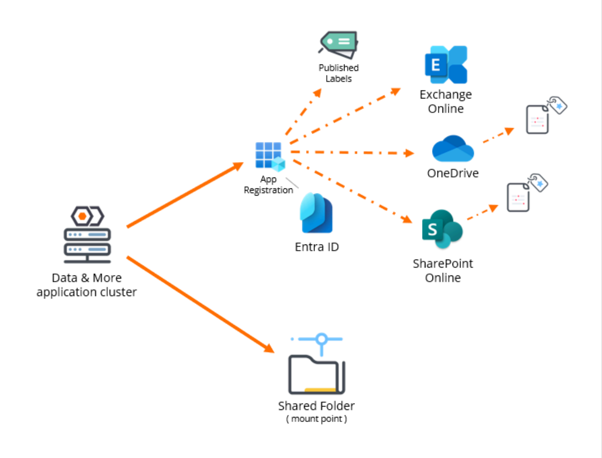
As you can see in the flow above, you don't have to use any manual classification or use the Purview Auto Classification feature. However, we recommend using Purview manual classification for classification types that require human insight to assess the necessary protection level. Manual classification is often not Privacy Classification per se (Privacy Classification, which can be automated) but rather business-related risk-based assessments such as confidentiality and IP.

It still makes sense to use Purview Auto Classification to apply sensitivity levels related to organizational-specific information types. This is especially useful if you already use EDM for other Azure services.

Data & More for Purview Implementation

The implementation of Data & More for Purview is straightforward and can be implemented via a simple app registration from the customer's perspective. For details about the app registration, please consult the D&M documentation here[6].

The app registrations connect the Data & More Application Cluster to the Organisations Microsoft 365 Tennant, and if you use DAMCE-PW as a SaaS - then the implementation is done. If you want DMSCS-PW on-premise, you can find the specifications for the application cluster here[7].



The DMSCS-PW will then begin classifying your organization's data in Microsoft 365. Within a couple of days, the DMSCS-PW will get an overview of the most common document classes (depending on tenant size), and each will be mapped to your organization's Sensitivity Labels.

This mapping is usually done in two one-hour workshops, and then the actual application of sensitivity labels can begin.

[6] https://support.dataandmore.com/en/knowledge/rights-overview

[7] https://support.dataandmore.com/en/knowledge/what-is-the-on-prem-client-cloud-requirements-for-enterprise-installations

Data & More Privacy Classification

Data & More's data classification team has spent over five years developing a comprehensive privacy compliance classification system that is used, tested, and quality assured on billions of data each day. This system can identify privacy data as required by law for each country. It is maintained and optimized daily by a multilingual team.

DMCS-PW provides an entire Data Privacy Classification in 28 languages, including local health information, unions, religious and sexual orientation, and all the sensitivity categories required in eg. the EU, North America, and Canada. Examples below:

Confidential Personal Data Document Classes (examples)

- European & international ID
- ID card, number, or information
- European & International Social Security info
- Social security card, number, or information
- European & international health cards
- Health card, number, or information
- European & international drivers' licenses
- The card, number, or information
- European & international passports
- The passport, number, or information
- Credit cards
- The credit card, number, or information
- Tax information
- Tax returns, etc.
- Residence permit
- Permits and or information in them
- Salary information
- Pay slips, etc.
- Employment documents
- Contracts etc.
- Recruitment (Application/job offer/CV)
- A wide range of information related to the recruitment process, job applications, CVs, job interviews, etc.
- Bonus agreements
- Dismissal or resignation
- Terminations or resignations
- Written warnings
- Expulsions

Criminal offenses document classes (examples)

- Criminal record
- Criminal records and information about them
- Offenses, fines, and convictions
- Convictions, fines, etc.

Sensitive personal data document classes (examples)

- Health info
- Diagnose
- Illnesses
- Medication
- Sick leave
- Health evaluation
- Prescriptions

Trade union membership (examples)

- Membership of a trade union

Orientations Belief & Origin (examples)

- Which country do you come from
- Ethnicity
- Membership in a political party
- Religious orientation
- Member of a religious church
- Religious congregation
- Gender types
- Information about sexual orientation

Non-sensitive personal data document classes (examples)

- Pictures with a face
- Used for classification in different document classes
- Travel information
- Travel bookings
- Reservations
- Check-ins, a.g.., showing where you have been at any given time

The above list is an example of some of the Privacy classifications. For each class, there are hundreds of thousands of classification elements that are used to identify positives and remove false positives.

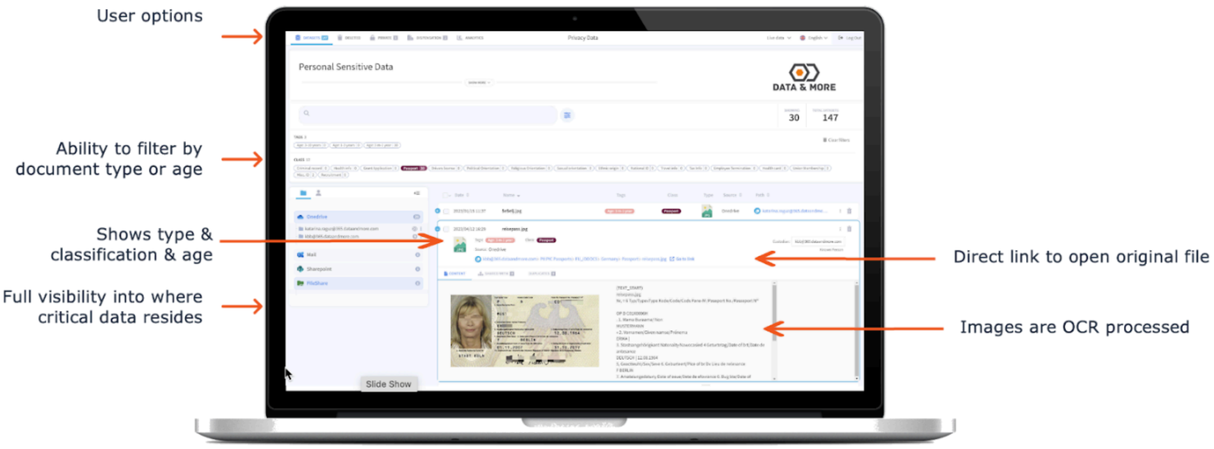
End-User Validation for Purview

Data & More End-user validation for Purview gives the data owners a complete overview of the data that has been auto-classified as non-compliant. The data owner can then use the report to mark documents as misclassified, private, or dispensation, depending on the individual's needs.

It gives the ensure quick preview of the and insight into:

- The total amount of data that must be cleaned up
- Document Age
- Document Class
- Document Location
- Document Type
- Document Owen
- Data Subjects

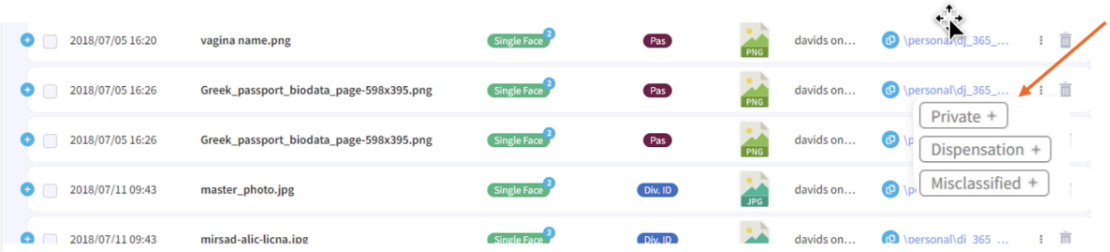
By combining the age filter with location and class, the end-user can very quickly identify, e.g., passports older than five years that are located in the inbox, or health certificates older than ten years in an HR SharePoint site.


User options
Ability to filter by document type or age
Shows type & classification & age
Full visibility into where critical data resides
Direct link to open original file
Images are OCR processed
Privacy Data
Personal Sensitive Data
DATA & MORE
30
147
Onedrive
Mail
Sharepoint
FileShare
Slide Show

(The End-user Validation Report gives each end-user easy access to their own Privacy Data)

One of the many advantages of the D&M for Purview End User Validation Report is that the end user can easily check the data and mark data that they need to keep with "Dispensation" or, in the case of Private data, as Private Data. This type of marking can be customized to specific organizational needs.

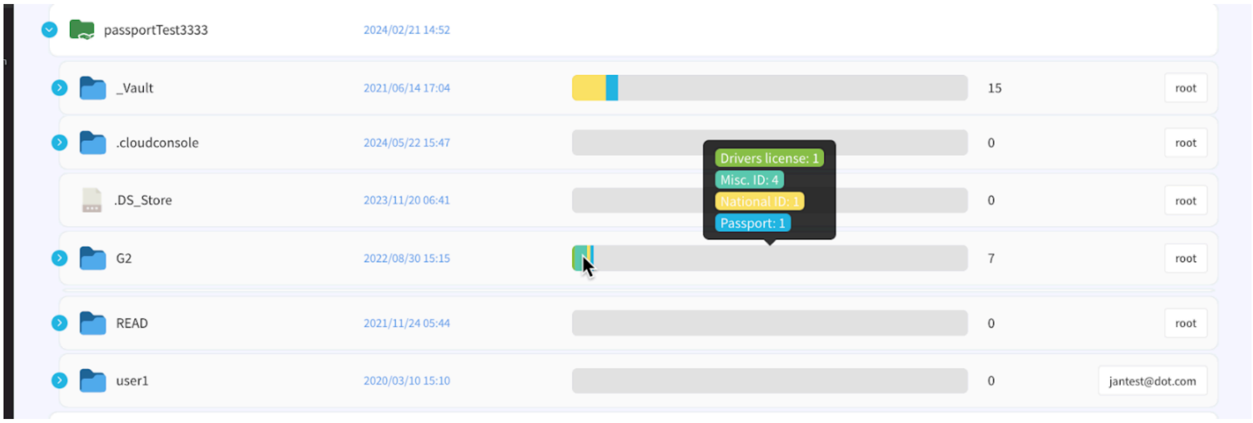
If the end-user thinks that a document has been falsely classified, the end-user can mark it as misclassified as input to the Data & More's classification team (the customer may decide what options are available the the ned-user).



2018/07/05 16:20	vagina name.png	Single Face 2	Pas	PNG	davids on...	\personal\dj_365_...
2018/07/05 16:26	Greek_passport_biodata_page-598x395.png	Single Face 2	Pas	PNG	davids on...	\personal\dj_365_...
2018/07/05 16:26	Greek_passport_biodata_page-598x395.png	Single Face 2	Pas	PNG	davids on...	\p
2018/07/11 09:43	master_photo.jpg	Single Face 2	Div. ID	JPG	davids on...	\p
2018/07/11 09:43	mirsad-alic-licna.jpg	Single Face 2	Div. ID		davids on...	\personal\dj_365_...

Custodian Mapping for Purview

Proper data management requires that one (or more) person be responsible for a specific data repository. This is easy for Mail and OneDrive, as it is the owner's responsibility. However, with SharePoint and FileShare, this is often much more blurry due to the rotation and churn of employees. Data & More offers a great tool to identify, manage, and maintain "Custodians." The best thing about custodian mapping is that it works by a risk-first approach, so you can start mapping the data that is a risk and leave the rest for later.

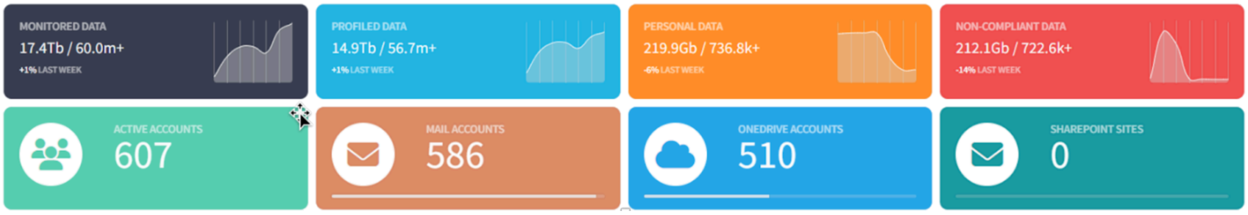


passportTest3333	2024/02/21 14:52		
_Vault	2021/06/14 17:04	15	root
.cloudconsole	2024/05/22 15:47	0	root
.DS_Store	2023/11/20 06:41	0	root
G2	2022/08/30 15:15	7	root
READ	2021/11/24 05:44	0	root
user1	2020/03/10 15:10	0	jantest@dot.com

Compliance analytics

Compliance analytics serves a dual purpose. First, it helps C-level, DPO, and middle managers follow up on any irregularities in the Data Privacy landscape. That could be data sources with a large amount of unmanaged Privacy Data or individual users who have marked large data sets with

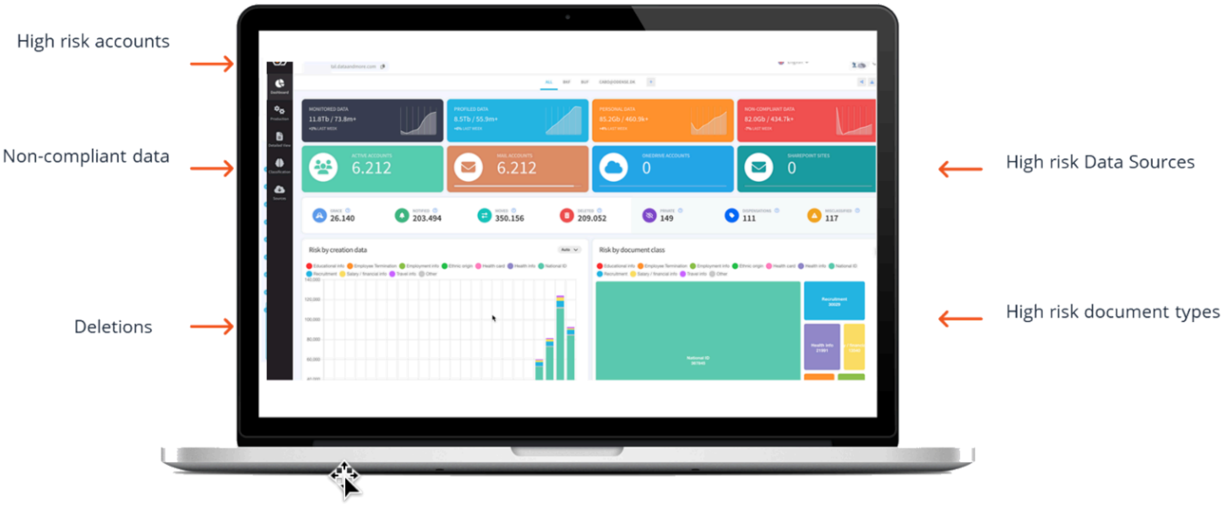
Dispensation or as Private. The second purpose of Compliance Analytics is to provide comprehensive and trustworthy documentation to auditors -both internal and the Data Privacy Authorities.



(Dashboard header in Compliance Analytics)

Key metrics in the Compliance analysis include:

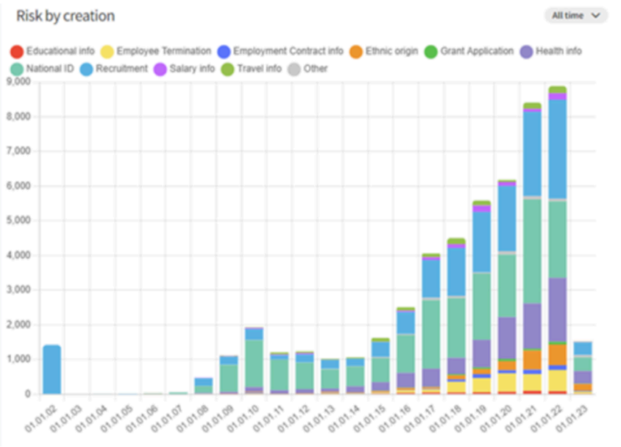
- Total number of Data Subjects
- Total number of datasets with Privacy Data
- Document age
- Data location
- Owner or custodian
- Document classes
- Data for data minimization (Owner, Type, Location)

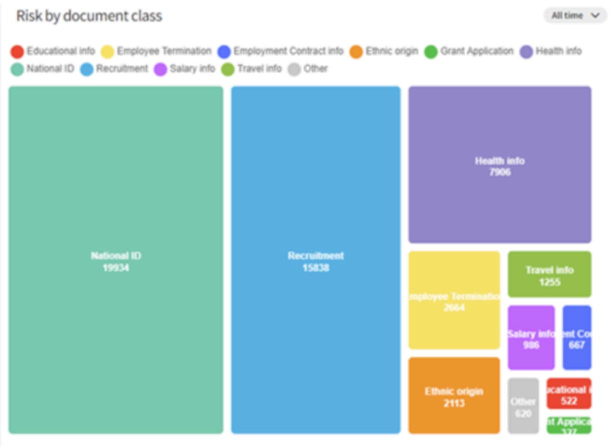


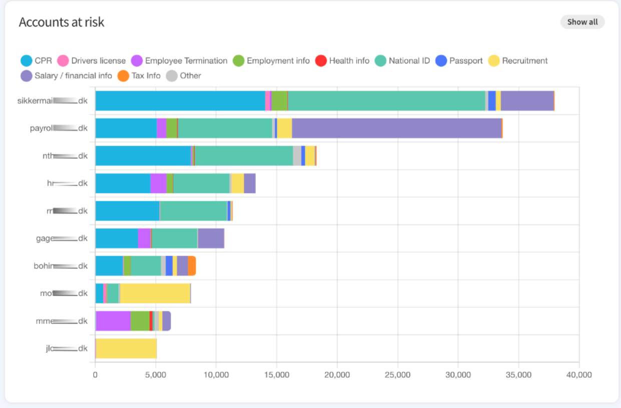
The above reporting can be built in Power BI, but DAMCE-PW can also report on:

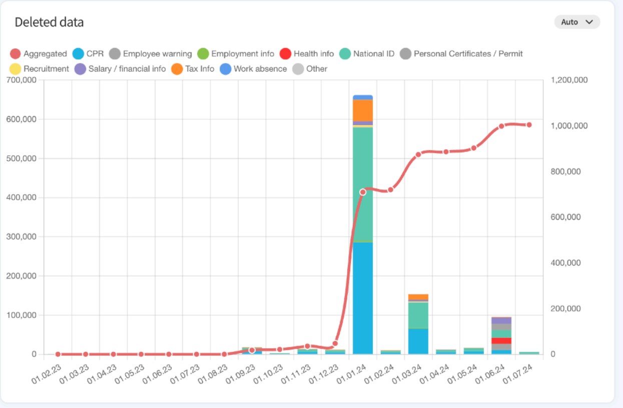
- Privacy Data that can't be tagged in Purview should have Logfies, HTML, and XML
- Privacy Data that is not connected to MS 365, such as old file-share and on-premise SharePoint
- Data in Exchange that have not been classified in transit

Here are two examples of Compliancy Analytics Graphs:









In the Data & More for Purview Compliance Analytics, the report can be filtered by AD Groups or specific policies or tags.

The report can be used to show specific document classes or company-wide but will never show the actual content data.

Label Review and Re-Labelling

For any organization relying on user labeling, having a level of oversight to ensure the right labels are being applied to the right documents is extremely valuable. Data & More for Purview can compare the contents of the document (i.e. the privacy scan results) to the label that's been applied to the document. Where they don't match, most likely Purview isn't properly protecting that document and the document label should be reviewed. Data & More for Purview can provide the reports needed for document review and, if you wish, re-label documents to the correct label.

Similarly, if your business requirements change and you need to change the existing labels on some (or all) of your documents for any reason, Data & More for Purview can automate that task and provide the reports you need before and after to confirm the re-labeling was completed as required.

Documents Not Protected by Purview

Many organizations are unclear on what document types provide support Purview labels, therefore can be protected by Purview policies. A list of the supported document types is available here[8]. Data & More for Purview will scan (almost) all text and image-based files, providing you with a complete picture of the privacy information you are storing in your unstructured data repositories.

This visibility allows organizations to quantify the size of this 'gap', report on the specific files that fall outside the scope of Purview protection and develop a plan for how to manage these files. Also, a reminder the End-User Validation, Custodian Mapping and Compliance Analytics provided by Data & More for Purview all work as expected on these documents.

—

[8] https://learn.microsoft.com/en-us/information-protection/develop/concept-supported-filetypes

07 Final Thoughts

There is nothing like a hands-on evaluation of a white paper. If you want a demo of our software, you can try demo.dataandmore.com. It will give you five users out of the box.

If you send an email stating that you would like to evaluate Data & More for Purview, we will provide you with 20 more users to test and a free compliance workshop if your organization has more than 200 employees.

Contact us via:

w: dataandmore.com // m: support@dataandmore.com // p: (+45) 4290 1070

—