Do you really know what sensitive data you have?

A new approach to identifying and managing sensitive data throughout its lifecycle

A Nogacom White Paper

Thomas Quednau

Introduction

Data environments are chaotic. Most companies have millions, if not billions, of files swirling around their organizations: documents, PDF files, emails, spreadsheets, presentations, etc. More and more of these files are created every day - many containing sensitive business information - by users across the organization, from the newest administrative assistant to top C-level executives. Furthermore, users are freely storing, sharing, emailing, copying and re-purposing the content of these files—often with little, if any, regard for the company's security and compliance best practices.

Given the volume and the continuous motion of the data, as well as the complexity of today's data environments, which often span multiple geographies and technologies, it's near impossible to get a handle on what business data actually exists and what its value is to the organization – let alone figure out whether it contains sensitive content or not.

This white paper outlines some of the key problems associated with trying to identify sensitive unstructured data. Nogacom contends, however, that identifying sensitive data is only part of the problem. Companies must also drill down to understand the underlying business practices and processes that created the existing chaotic and unsecure data environment. This is so that they can find and plug the holes in the organization and fix faulty business practices and processes and ultimately create an organized, safer and more effective environment for their crown jewels – their sensitive data. Another aspect of the problem that is often left by the wayside is the fact that the nature of data can change over time – either because of changes to the document or because of changes to the business: one day a document may be sensitive, and the next its not. Nogacom believes that all products and processes implemented by companies to manage their sensitive data must also addresses the potentially changing nature of sensitive data throughout its lifecycle.

Lastly, NogaLogic presents its solution and outlines its own best practices approach to identifying and governing sensitive data throughout its lifecycle.

The problems associated with identifying sensitive data

Companies don't know what sensitive data is

Most companies know that sensitive data is data that includes information on the company's financials, business/product/marketing/sales plans etc, as well as employee and customer information. However the complexity and scope of sensitive data is actually far wider and many companies don't know the full extent of what it really covers.

The multitude of regulations, laws and company policies governing data add a wide range of data – and complexity – to the long list of what should be considered sensitive data. For example, they include PII data (Personally Identifiable Information); documents that could potentially be related to a lawsuit – even if the litigation process has not yet been initiated; documents containing two or more pieces of information which separately are not sensitive, but together in the same document are, etc. In addition, the sensitive information in a document may not be immediately visible in the content of a document – it may be contained in its properties or in an image embedded into the document.

Nogacom's own research has identified the following types of sensitive data:

1. Information related to the running of the business, such as a business plan, financial information, commercial details, customers, employee information etc.
2. Two or more pieces of information which separately are not sensitive, but together in the same document, they are. For example, the combination of a person associated with a particular company.
3. Documents containing specific sensitive regular expressions, such as credit card numbers, patient record numbers, bank account information, ID numbers etc., and documents containing Personally Identifiable Information (PII).
4. Documents tagged as 'confidential/top secret' etc.
5. Information deemed sensitive by regulations.
6. Information needed for an eDiscovery process - information which on the face of it, may not seem sensitive at all, but in the context of the litigation, is.
7. Certain data that is deemed to be a business record, which companies are required to retain.
8. Documents that are sensitive based on certain meta data, such as a specific author.
9. Documents related to a specific topic which management has deemed sensitive.

Sensitive data comes in many varieties. It is critical to correctly identify it in order to manage and protect it appropriately. To err on the side of caution and classify any document which may possibly be sensitive as sensitive is not good business practice. These documents will be subjected to unnecessary restrictive policies that will prevent their legitimate and necessary use by other employees. This will harm the business and cost the company in terms of unnecessary management overhead, storage and protection software solutions.

Companies don't understand the changing status of a document throughout its lifecycle

Another aspect of the problem of identifying sensitive data is that the identity and value of data may change throughout its lifecycle – from creation through to destruction - depending on changing business needs, M&A activity, new regulatory mandates, updated content, etc. For example, company A acquires company B, which was a competitor. In the past, information about company B was identified as competitive and sensitive, but following the acquisition Company B is no longer a competitor and most of the data about it is no longer sensitive. In this case, the sensitivity and business value of documents related

to company B have changed over the course of the documents' lifecycle, and therefore they will need to be managed differently during at each stage.

Companies don't know what their sensitive data does

Identifying sensitive content is only part of the picture. Companies also need to understand how this data is actually used across their organizations – where it's stored, what versions and copies of this data were created, to whom sensitive data is being distributed, what access permissions exist for this data, and who has actually been accessing this data. This analysis is crucial to understanding what underlying processes are occurring throughout the business which facilitated the current chaotic data environment, and why they are occurring. Maybe access permissions have not been reviewed and updated recently; maybe new employees have not been briefed on company policies regarding the handling of sensitive data; maybe new technologies are needed to help prevent data leaks. Without this dimension of insight, the process of identifying and managing sensitive data will only be partially successful and the data chaos will persist.

Take, for example, a person who pastes part of a sensitive document into a new document, and then emails it to a colleague who saves it, without restricting access to it, in his personal directory on the network. Without insight into all the processes that occurred during this document's lifecycle – identifying this document as a version of a sensitive document, knowing who emailed it to whom, where its stored and what access permissions exist – this document would never have been identified as sensitive, and all the improper processes that occurred during the handing of this document would never have been discovered.

Companies believe that DLP solutions will identify their sensitive data

The size and the complexity of their data environments together with the lack of insight into what sensitive data they have, and what they need to do in order to identify and then govern it, scares companies. They simply don't know where to start. Some try quick-fix solutions, some implement technologies that focus on a specific aspect of the problem; some rely on their employees, while others rely on consultants to do the job. Many don't attempt to tackle it until its sensitive data has been compromised. And there are many companies who believe that their Data Leak Prevention (DLP) product will do this job.

While DLP solutions utilize a variety of techniques to discover sensitive data there are a number of problems using them to identify sensitive data. First, DLP products are designed to prevent data leaks rather than identify sensitive data and support good information governance. Second, while DLP products do have a wide range of data discovery capabilities, they do not actually identify it. They simply discover its existence but do not provide any information on the actual identity of this data or its business context and value to the specific organization. Third, DLP products do not provide any insight into how and why sensitive data is used and distributed across the organization, even though this insight is necessary for defining relevant data protection polices as well as fixing or updating flawed business practices that cause data leaks.

Identifying sensitive data using NogaLogic

Nogacom provides companies with a simple, flexible yet extremely powerful, comprehensive and automated solution for identifying and analyzing sensitive data throughout its lifecycle.

The core foundation of Nogacom's NogaLogic solution for information governance is its unique ability to automatically identify and classify business data – particularly sensitive data - based on its business context and value to each specific company.

The business context of a document objectively identifies the essence of a document and its business value to the organization, and differentiates it from others. And it impacts everything to do with how this document is managed and governed, from its access permissions and distribution to its storage location and retention requirements.

Take, for example, two spreadsheets, one containing a list of key sales people, their top customers and their revenue targets, and another containing a list of all the employees in the R&D department and the amount they each contributed to the another staff member's birthday present. On the face of it, both documents look similar: each contains a list of employees against dollar amounts, but in reality the business context of each these documents will identify and distinguish one from the other, and determine how each one is managed and protected.

Clearly the document containing customer names and sales targets is a sensitive document that needs to be governed and protected according to appropriate regulations and internal business best practices, while the second is pretty innocuous and does not require such stringent security protections. But without being able to identify these documents based on their business value to the organization it is likely that either neither will be effectively protected – which would put the company at risk - or both will be over protected which will waste storage space (and costs) and require unnecessary management overhead. Furthermore by placing unnecessary restrictions on its documents, companies will likely harm the flow of business information across the organization and employees will more likely try to find ways to bypass these restrictions which will, in turn, invalidate the data security measures in place and put the company at risk.

Identifies all sensitive documents

NogaLogic doesn't just identify documents by their business context, it can also specifically identify sensitive content within documents, or various combinations of information which together make a document sensitive. These include:

- Credit card numbers, regular expressions, pattern matching and PII data
- Pre-defined regulatory templates that identify data deemed sensitive by regulations
- Documents tagged with various sensitive classifications such as "top secret", "confidential" etc.
- Information needed for an eDiscovery process
- Certain business records which companies are required to retain
- Sensitive information contained in meta data/properties tags

Provides a full picture of the data lifecycle

Once it's identified the data, NogaLogic automatically shows where each document is stored and how it's being used across the organization, giving companies full visibility into their data's lifecycle and helping them understand the underlying the business practices that created the existing data environment. Data lifecycle details include:

- **Copies, Versions, Related Content.** NogaLogic automatically identifies all copies and versions of any document or sensitive content that has been cut and pasted into a new document —regardless of where it is located, its format or how it is named. It also shows the date and author of any revisions and the actual changes that were made in the text.
- **Distribution**. NogaLogic automatically tracks the distribution of documents by email, both inside and outside the company.
- **Audit History**. NogaLogic provides a full audit history on each document, showing who has accessed a document and made changes, when these changes were made, where each new version of a document was stored, and other actions performed on a file.
- **Access Permissions.** NogaLogic shows who currently has access permissions on documents.
- **Location identification**. NogaLogic identifies the storage location of each document.

NogaLogic also provides sophisticated search capabilities which enable the user to fine tune data selections based on a wide variety of parameters. Furthermore, users can organize specific data selections into user-defined categories (Views).

NogaLogic's analysis is presented through a UI and in detailed, drill down reports as well as a dashboard which shows the current status of sensitive data across the organization and enables users to easily identify sensitive data which is at risk.

NogaLogic also includes a sophisticated and highly flexible policy management engine through which users can move, copy, and tag documents selected based on their business context and a wide variety of parameters.

Supports the changing nature of a document

NogaLogic is designed to handle the changing nature of a document throughout its lifecycle. NogaLogic automatically classifies new data as it enters the data environment and associates that data with existing data that matches that business context. It also continually reviews existing data for any changes that impact the business context of the document or the data lifecycle, and updates its own database, data views and reports accordingly. If a certain business topic becomes sensitive to the organization, the user simply changes the status of this business topic within NogaLogic accordingly. All documents related to this business topic are then automatically re-classified as sensitive.

Best practices for identifying and governing sensitive data throughout its lifecycle

Nogacom has developed a best practices process for identifying and managing sensitive data based on its NogaLogic solution. This step-by-step approach is broken down into iterative stages that make the process easy, intuitive and manageable. These stages include:

1) Identifying and classifying all data based on its business context
2) Analyzing, organizing and cleaning up all data
3) Identifying and organizing sensitive data
4) Discovering, analyzing and resolving problems

Identifying and classifying all data based on its business context

The key to successfully identifying and analyzing sensitive data is to first classify all business data by its business context. This initial identification provides the foundation for analyzing all data within the context of its meaning and value to the specific organization.

Analyzing, organizing and cleaning up all data

Given the extent of the chaos that exists within most data environments today, its highly recommended to first perform an initial clean up the entire environment before attempting to identify and analyze the sensitive data. This includes getting rid of personal files, unnecessary copies, large, space hogging audio or video files, archiving old documents and removing documents that have passed their data retention expiration dates, etc.

NogaLogic intuitively guides users through this process via its dashboard, which provides an up-to-date snapshot of all data, enabling users to easily see key problem areas within their data environments. Once these have been identified, NogaLogic provides detailed reports that hone in on the source and scope of the problems, and a comprehensive and flexible UI which serves as a workspace where users can create their clean up action plans. Next, users can use NogaLogic's policy manager, third party solutions and manual actions to implement these action plans. At the end of this stage, companies will have made significant progress in controlling their data chaos.

Identifying and organizing sensitive data

Now that the data environment is more manageable, it's possible to identify and organize the sensitive data within it. Through NogaLogic's UI and using its highly flexible search and correlation capabilities each type of sensitive data is easily identified and can be tagged and/or organized into relevant user-defined categories ('Views'), based on the company's specific business and governance needs.

Once Views have been defined, all new or updated data that match the View's criteria will be automatically added to that View. Furthermore, Views are dynamic and can be adjusted to meet changing business or governance needs. Users can update a View's criteria at any time, and the list of documents included within the View will be updated accordingly.

These Views can be exported to data protection solutions such as DLP, thereby providing these technologies with the missing 'business context' perspective to add to their own data discovery processes.

Discovering, analyzing and resolving problems

The last stage is an iterative one. Users should first check the dashboard to identify the most vulnerable and problematic Views, such as those with sensitive documents which are dangerously exposed, or those with sensitive documents stored in unprotected data repositories. These are obviously the documents that the company should focus on first.

Using NogaLogic's reports users should then drill down to fully analyze how these sensitive documents are handed across their organizations including what access rights exist for these documents and whether they are appropriate, who actually accessed and made changes to sensitive documents, whether these documents were emailed, and to whom, where they are stored, and whether copies and versions of these documents exist, and where they are stored.

With this insight, users can then take the necessary actions to effectively protect and govern their sensitive data data. These can include moving documents to a secure data repository, adjusting access rights, ensuring that copies and versions of sensitive documents are also protected, implementing 3rd party technologies, changing business practices or educating users in the company's best practices for handling sensitive data.

Once the most critical at-risk data has been appropriately managed, users should go back to the dashboard to prioritize the 'Views' to work on next, based on the level of risk and company priorities.

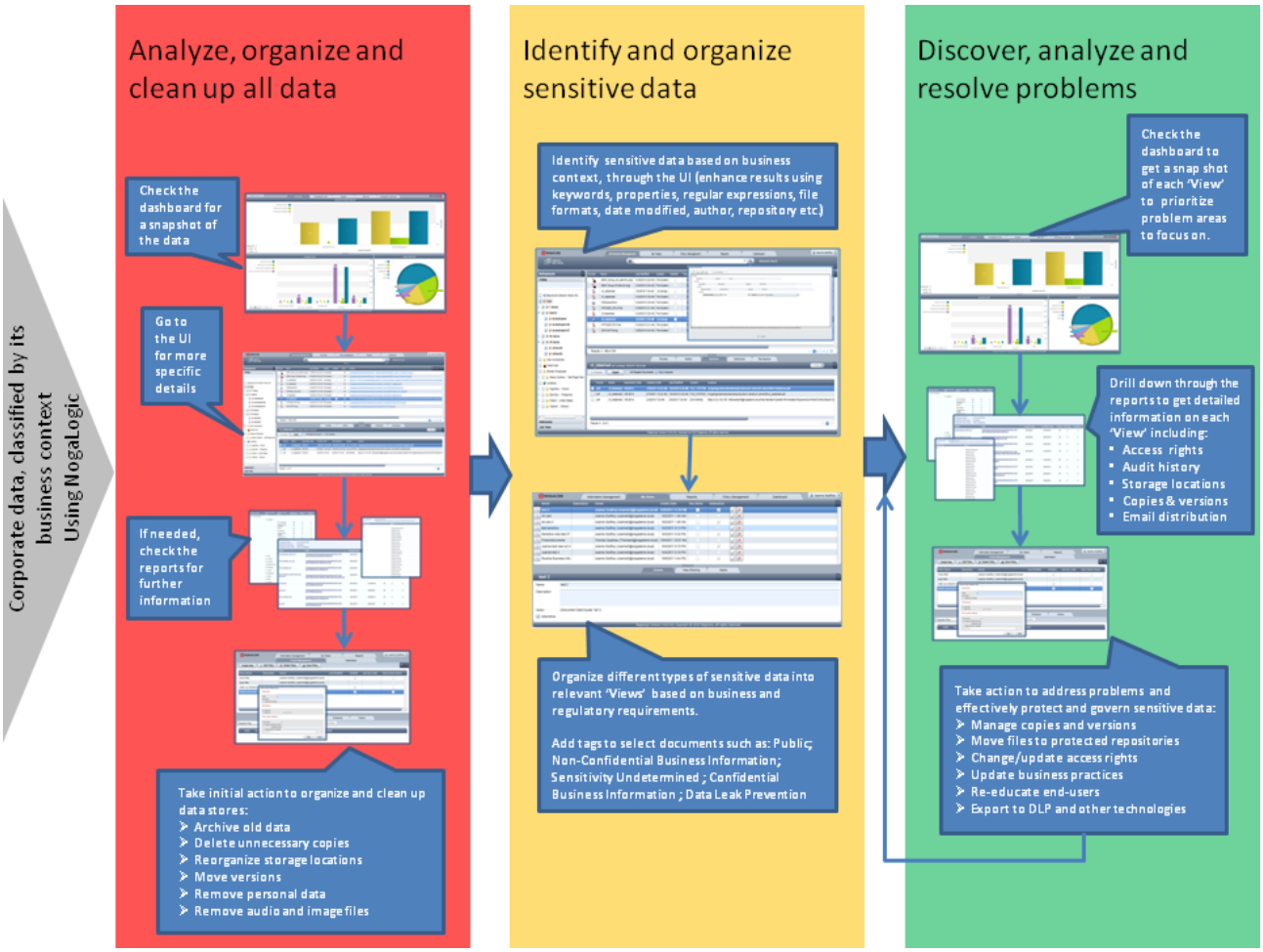

Figure 1: Nogacom's Best Practices Approach for Identifying and Managing Sensitive Data

Conclusion

Identifying, managing and governing sensitive data is critical for the success of the business as well for regulatory compliance, but its also a complex and multi faceted process. Many companies facing this challenge don't know where to start, what they really need to do, and how to go about actually doing it.

Nogacom gives companies this critical starting point. It delivers a powerful and meaningful way to identify, analyze and manage sensitive data, together with a structured methodology that simplifies the process and makes it manageable and effective.

The data environment and the business are both evolving organisms, and they can both change on a daily and sometimes even hourly basis. NogaLogic is dynamic and flexible and can automatically adapt to these changes to help companies ensure that their data environments are in sync with their business needs, at any given point in time.

Ultimately, Nogacom helps companies significantly reduce their risk exposure by enabling them to take control of their sensitive data and eliminate much of the chaos within their data environments.

About Nogacom

Nogacom delivers software and services offerings for information governance that give companies the power to make their unstructured data secure and compliant. With NogaLogic users can quickly easily, accurately and cost-effectively analyze their unstructured data – particularly sensitive data - and identify and address risks and problems.

NogaLogic solutions support a wide variety of use cases including data risk assessments, support for Governance Risk and Compliance (GRC), PCI DSS compliance, Cloud Computing, eDiscovery, data migration and archiving, Electronic Data Records Management, and much more. Nogacom's customers include leading financial institutions, technology and telecommunications, government agencies, transportation, and waste management companies.

Nogacom headquartered in Wadgassen and Hamburg Germany, Zurich and Locarno in Switzerland and Leiden in The Netherlands. The company is also supported through a network of resellers. More information on Nogacom is at www.nogacom.eu