Classify First, plan earlier

A new approach to effective data governance through data classification

A Nogacom White Paper

Introduction

Information is an organization's most critical asset. Therefore most companies are required to adhere to various internal and external regulations that govern the way their data should be managed in order to protect customers, shareholders and investors, partners, employees and anyone involved with the organization.

Corporate data environments, however, are usually pretty chaotic. Most companies have a huge number of files containing unstructured data scattered all across their IT environments: documents, PDF files, emails, spreadsheets, presentations, etc. More and more of these files are created every day by users across the organization, from the newest administrative intern to top C-level executives. Users are also freely storing, sharing, copying and re-purposing the content of these files—often with little, if any, regard for security and compliance best practices. This chaos directly translates into lost revenue, lost business opportunities and lost productivity, and exposes the company to regulatory compliance failures together with significant financial and legal liabilities.

Data classification – the process of identifying and categorizing data - is the only way to take control of the data environment so that you can then effectively govern it. The conventional approach to data classification dictates that you first plan your data governance strategy and then implement it with appropriate data classification schemas and technologies that address your specific governance challenges and needs. But in reality this approach has proven to be fundamentally flawed, difficult to apply and ineffective, as companies simply have no idea how and where to begin. As a result companies still find themselves exposed to the significant risks and liabilities of compliance failures.

This white paper describes Nogacom's new, best-practices approach whereby the steps are reversed: first you classify the data using Nogacom's NogaLogic information governance solution. With the intelligence gleaned from the data classification process, you then have the necessary insight to plan and implement relevant and effective data governance strategies.

This approach eliminates the significant challenges, complexity and fear that companies have of implementing a data classification process, and helps avoid unnecessary spending on technologies, infrastructure and consultants.

Data classification is the Key to Effective Data Governance

The sheer volume of data dispersed across the organization, the wide variety of content, and the complexity involved make it impossible to manage and govern the data environment manually. The solution is automated data classification which is a software-driven process that automatically identifies and categorizes data so that it can then be governed. However, most companies have no idea how to even begin to plan an automated data classification process and are often deterred by the fear of the significant preparatory work involved, such as deciding data classifications schemas and data governance strategies that need to be embedded into the process. They are also concerned by the perceived extensive manual efforts, such as figuring out keywords to track, as well as the extent of end-user involvement that will be needed during the classification process.

Nogacom contends that companies can only effectively safeguard and govern their data when they first know what data they actually have, where it is and what is being done with it. Without this insight, organizations are flying blind—and they can't figure out and implement the data governance strategies they need to protect their data and limit their exposure to financial, legal and business losses, and other consequences resulting from lack of information governance.

Nogacom therefore presents an alternative to the conventional approach to data classification that dictates that you first plan and then classify in order to govern your data. The Nogacom approach suggests that it's more effective to classify first and then plan your strategy, because it's only once your data is classified that you can really uncover what's going on with your data environment. And only then can you begin to make intelligent decisions on how to govern and protect it.

How do I classify data without a plan?

Nogacom's NogaLogic is a data classification solution for information governance that automatically knows how to identify and classify your specific data. It is fully automated, and does not require any advance planning or manual work. So there is no need to develop a classification methodology in advance or a list of keywords, and there's no end-user involvement in the data classification process.

Through its data classification process, NogaLogic provides unparalleled visibility into unstructured data. It identifies, classifies and organizes unstructured data based on its *business context and value to your specific company* – and it tells you where this data is stored and how it's being used across the organization.

At the end of the classification process all documents are automatically and accurately classified and organized according their actual business context—whether this context is related to a customer, a patient, a product, marketing program, a sales channel or supplier, an office or any combination thereof – without the need for any pre-conceptions, preliminary or manual work or end-user involvement.

What is this "business context" and why is it important?

Every piece of unstructured business data is about something that relates to the organization. It may be something related to a customer, a patient's details, a product, marketing program, the financial performance of a business unit, a sales channel or supplier, an office location, or any combination of such things. This is its "business context."

Obviously, the business context is key to governing a document, because the business context is what differentiates a proposal for a million-dollar deal from an invitation to the company party. Furthermore,

the business context of a document affects everything about how this document is managed and governed, from its access and distribution permissions, to its storage location and retention requirements.

Classifying documents by their business context is far more meaningful to the organization than simply tagging documents with somewhat arbitrary tags such as “confidential”, “top secret” or “public” because it accurately identifies the document in terms of the business issues contained within it.

It’s also far more accurate to classify documents based on their business context than to identify documents using keywords. This is because keywords require you to know the exact keyword to search for, yet in reality a keyword may appear in many different forms within documents or may not appear at all. For example, a particular business issue may be referred to using partial information or a subset of information related to that topic, such a company referred to by an acronym rather than its full business name (G.M. instead of General Motors) or a person referred to by a an email address or his or her ID number rather than his or her name. If you don’t explicitly search for G.M., you won’t be able to identify those documents that refer to the company as just G.M.

To ensure accuracy and completeness, a data classification technology requires a deep understanding of complex texts, which were written by humans for humans. Humans, for example, can easily distinguish between the meaning of the string "nice" in the sentences "It was a nice day today”, “He works for NICE Systems", "The event was held in Nice, France", or "We met with David Nice." Ordinary keyword-based text retrieval systems cannot make these distinctions.

To accommodate the various semantic variations that appear in the unstructured text and to ensure accuracy and completeness, NogaLogic utilizes various morphological, lexical and semantic Natural Language Processing (NLP) techniques during its data classification process.

Using these techniques the various business-related entities referenced in documents are accurately identified not only by their actual names, but also by other properties—such as email addresses, phone numbers, or other unique or semi-unique identifiers. For example, a person whose e-mail address is “lazer@sandiego.edu” will be associated with the University of San Diego since "sandiego.edu" is that University’s internet domain. By using these properties, documents will be classified with much higher recall than when classifying based on name search only.

Furthermore, these technologies enable NogaLogic, for example, to distinguish between the color brown and the last name “Brown” in a sentence. In the example mentioned above, NogaLogic knows to accurately identify and differentiate between the different ‘nices’ - the company NICE, the place Nice, the Person Nice and its use as a simple adjective in the sentence “have a nice day”.

What about documents containing sensitive business information?

Another key advantage of NogaLogic’s business-context data classification is that it can easily identify documents containing sensitive business content. Documents are generally considered sensitive in the following three situations:

- They contain information that is confidential and sensitive to the business, such as financial information, information on a new product or marketing campaign etc.
- They contain information needed for an e-Discovery process and/or Electronic Data Records Management (EDRM).
- They contain information deemed sensitive by the various regulations.

NogaLogic's business-context data classification can automatically identify all these types of sensitive documents – whether the sensitive content appears as a single occurrence in a document, or whether there are multiple types and/or occurrences of sensitive information in a single document.

How else does NogaLogic help analyze unstructured data?

NogaLogic doesn't just classify data by its business context. NogaLogic's performs many other important and unique functions during its classification process which significantly enhance the value of the classification and enables companies to gain full visibility into the behavior of their unstructured data and their end-users throughout the data's lifecycle. Some of these features include[1]:

- **Version tracking**. NogaLogic automatically identifies all copies and versions of any document in any file format—regardless of where it is located or how it is named—as well as the date and author of those revisions and the specific changes that were made in the text.
- **Distribution tracking**. NogaLogic automatically tracks the distribution of documents by email, both inside and outside the company.
- **Document auditing**. NogaLogic provides a full audit history on each document, showing who has accessed the 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.
- **Credit card number identification and regular expression pattern matching**. NogaLogic identifies credit card information and regular expressions (such as social security numbers, bank account numbers, healthcare provider numbers etc.) in documents and text fields.
- **Advanced filtering and querying.** NogaLogic includes the ability to filter results by file format, location, date and type as well as create complex search queries using a wide range of parameters and document properties.
- **Location identification**. NogaLogic identifies the storage location of each document.
- **Centralized policy management.** NogaLogic includes a sophisticated yet highly flexible policy management engine through which users can move, copy, tag and share documents selected based on their business context as well as a wide variety of parameters.

[1] Detailed information on these features can be found on Nogacom's website at www.nogacom.eu

Assess and Plan an Effective Data Governance Strategy

NogaLogic provides an up-to-date picture of your data environment, organized by its business context and value to your organization. This now gives you the necessary intelligence to begin the data governance process: accurately assess your current situation, and plan the right strategies and resources needed to effectively govern your data.

Knowing the risks is half way to solving them. One of the first things you should do once you have a picture of your data environment is to assess your risk. With NogaLogic you immediately get a detailed understanding of what's really going on in your data environment – what documents you actually have (including copies and versions), where they are stored, who has access to them, and to whom they were distributed. With this information you can determine your immediate and long-term risks - legal risks, business risks, compliance risks - and then begin to figure out and prioritize the actions needed to mitigate them.

Knowing the past is the key to the future. To begin planning an effective data governance strategy you also need to understand the way employees currently use business data—the reality of how and why your data environment ended up the way it is today. Why were files stored in a particular place (against company policy)? Who is making new copies and versions, and why? Who is accessing files containing sensitive business information, and why? Which sensitive business data is being emailed, by whom, to whom, and why? etc.

NogaLogic provides deep analysis on how business data is currently being used so that you can then ask the right questions of employees in order to understand the 'how and why' behind your current data environment. Once you have the answers you can then begin to intelligently define and implement appropriate policies and best practices – as well as re-educate users – that manage and govern unstructured data throughout its lifecycle—and thereby bring essential order and accountability to an otherwise chaotic data environment.

Knowing the data is the key to protecting it. If you don't know what data you have, you simply cannot effectively manage and safeguard company-sensitive information. NogaLogic's insight is especially important for documents containing sensitive business information, and documents containing information that is governed by regulations, such as Sarbanes Oxley (SOX), Basel II, ISO 27001, Gramm-Leach-Bliley Act (GLBA), Health Insurance Portability and Accountability Act (HIPAA), Health Information Technology for Economic and Clinical Health Act (HITECH) Act, Federal Information Security Management Act (FISMA) PCI DSS, ISO 15489 (Records Management), the various Data Protection Acts, and many others.

It is important to note, that the sensitivity of a document may change throughout its lifecycle depending on changing business circumstances and/or updated content. For example, a company launches a new product onto the market. In the past, information on this product was sensitive, but now that it's publicly available it's no longer sensitive. In this case, the document will need to be handled differently during each stage of its lifecycle depending on its current sensitivity. NogaLogic is a dynamic solution that enables you to address changing business circumstances. NogaLogic can continually reassess the content of a document and can notify you when the content changes or other actions are taken on the data so that you can then adjust the way the document is handled and governed.

Furthermore, by using NogaLogic companies can avoid over-protecting documents that do not contain sensitive information so that they can be more effectively leveraged by users inside and outside the company. This improved utilization of information will, in turn, lead to better information governance and business performance, customer satisfaction, and other competitive advantages.

Knowing the data and who really is accessing it. Access rights to documents are a key part of effective data governance and one of the biggest concerns for companies. This issue has recently been in the headlines when celebrities' medical records were viewed by unauthorized personnel and then leaked to the media, in direct violation of HIPAA regulations.

NogaLogic shows who has access rights to the data (and which groups they belong to) and who is actually accessing the data. Since this information is correlated with the business context of the data NogaLogic provides a full picture upon which you can then determine appropriate access rights as well as identify inappropriate access.

Additionally, with NogaLogic it's easy to see if someone has access to data that he or she should not have, such as someone in the marketing department who has access rights to financial documents, or conversely, whether someone should have access to certain data but doesn't. With this insight you can fully analyze and then update access rights to documents to address information governance and legal requirements as well as business needs.

Knowing the data is the key to finding credit cardholder information in the data haystack. Compliance with the PCI DSS standard is mandatory for all companies handling credit card transactions. Not only can NogaLogic immediately identify all documents containing credit cardholder information, but it also shows where these documents are stored, who has access to them and who has accessed them, and to whom they were distributed via email. With this insight you can immediately rectify any PCI DSS compliance failures by, for example, changing current business practices, re-educating staff, adding security controls and changing access rights, and moving these documents to a secure repository (which can be done automatically through NogaLogic's policy management).

As PCI DSS compliance is an ongoing process, NogaLogic can be used to monitor documents containing card holder information that enter the data environment and alert on any PCI DSS compliance related failures and breaches.

Knowing the data is the key to categorizing it. NogaLogic gives you the critical insight needed to understand the essence of the data so that you can then determine meaningful and relevant designations for that data. For example you can now easily determine which documents should be categorized as "Confidential – Restricted Access", "Public Information – All Employees", "Limited to Executive Management Only", or "Sensitive M&A Information" etc. With NogaLogic you can even add the appropriate designation as a property tag to all relevant documents that meet the required criteria. This can be done automatically through NogaLogic's policy management.

The business is an ever-changing organism, and NogaLogic automatically updates the picture of your data environment hand in hand with your changing business. Therefore, if/when a change in the business warrants a change in the designation schema, it can again easily be done through NogaLogic's policy management.

Knowing the data is the key to being ready for eDiscovery. In an increasingly litigious society, your company must be prepared for an eDiscovery request at any time. It can be prohibitively expensive and enormously disruptive to attempt to find all the documents relevant to a specific lawsuit after a request has been received. And in some cases, it may actually be impossible to fully comply with eDiscovery request within a tight legal deadline. As a result, your company may be exposed to significant legal risks and/or may wind up going to court without a key piece of exculpatory evidence. Your company may also be fined for being unable to comply with the court's request. Even worse, it may find itself in a position where it has to settle a case that it could have won it on its merits.

Furthermore, the Federal Rules of Civil Procedure (FRCP) now require organizations to preserve relevant information when it learns, or reasonably should have learnt of pending or threatened litigation, or of a regulatory investigation, even before a law suit is actually filed. In practice this means that unless your organization prepares for such scenarios in advance, it may not be able to meet eDiscovery requirements without a major disruption to its operations.

With NogaLogic you are on your way to being prepared for eDiscovery. By automatically identifying and classifying all documents in advance, NogaLogic eliminates the need to hunt for documents once an eDiscovery process starts, is anticipated or is threatened. Furthermore, NogaLogic's advanced linguistic capabilities ensure that you find critical documents even if they use abbreviations, nicknames, or other potential semantic idiosyncrasies**.** Once you have identified the data you need for eDiscovery, you can then take the necessary steps to fulfill 'legal hold' obligations.

Furthermore, through NogaLogic you can automatically flag any new or updated documents that could potentially relate to the eDiscovery process and apply appropriate policies that preserve the integrity of these documents and ensure that users don't accidentally expose the company to new risks.

NogaLogic can also provide an audit trail for each piece of data—including date created/changed, author, and storage location—so you can document your compliance with court-mandated discovery requirements.

Knowing your data is the key to cleaning your data. Data cleanup activities are a key part of good data governance, and are essential when migrating data to a new data repository, consolidating data repositories or merging external data repositories into the company's existing data environment (such as the result of M&A activity). A clean data repository also helps optimize ETL performance and saves the company from wasting money and resources on managing and storing unnecessary data.

In addition to identifying documents by their business context NogaLogic identifies file formats and the document's storage locations. With this insight it's now easy to see what's really going on in the data environment. For example, you can see whether employees have stored personal audio or video files, photos, documents, or inappropriate content on the company's network. In addition, with NogaLogic it's immediately possible to see all the copies and versions of the same document, as well as content that has exceeded its required retention period, or is simply unimportant business documents.

This insight is crucial when developing data clean up strategies that require expunging personal or inappropriate material, deleting unnecessary documents and/or copies and versions, and archiving documents based on data retention requirements.

And once you've figured out what you need to do, NogaLogic can help you do it. Through NogaLogic's policy management you can define policies that will ensure that your migration process is efficient, complete, and supports your organization's best practices. For example, you can define a policy that automatically archives unnecessary versions to tape, or removes personal audio and video files.

Knowing the data makes cloud computing a realistic vision. Cloud computing is the current "next big thing". Everyone's talking about it, but they aren't all rushing to do it. Among the many fears regarding cloud computing is the very real concern regarding data protection and data governance failures.

One of the strategies that allow companies to take advantage of the performance, automation, flexibility, cost and resources savings of cloud computing environments without compromising on data protection and governance requirements is to move non-sensitive business information into the cloud but keep the

sensitive data on the ground. The obvious problem encountered when trying to implement this is identifying data and separating it out into sensitive and non sensitive information.

By identifying and classifying data, including sensitive data, based on its business context NogaLogic makes it really easy to see what you actually have and what you can move into the cloud. Furthermore, a cloud data migration project is also a good opportunity to clean up your data, since you probably don't want to spend time, money and resources managing employee's personal data in the cloud.

Knowing the data drives an efficient, cost effective and performance optimized storage strategy. With a clear understanding of your data you can now optimize your storage strategies. For example you can move unimportant data to lower cost storage repositories , important business data to higher performing storage systems, and sensitive data to appropriately protected repositories (as required by most regulations).

Furthermore many regulations require companies to retain and/or archive certain documents for specific periods of time, and dispose of them after the retention period has expired. NogaLogic provides all the necessary information to determine what needs to be retained, archived and disposed of, at any given moment in time. And through NogaLogic's policy management you can automatically create policies that implement these strategies across all data.

Knowing the data is the key to determining what resources you really need to get the job done. Having a clear picture of the current data environment and its risks, is invaluable in terms of planning what you need to do and what technology solutions you are likely to need to fulfill your objectives. It will also help you figure out timelines and manpower needs. Without this information you'll likely be guessing which will be very costly to you and your business.

Knowing the day is the key to getting more value from other security investments. When you have full visibility into the business value of your unstructured data, you can make more informed decisions regarding other data-related technology solutions – before and after purchase - such as Data Leak Prevention (DLP), Business Intelligence (BI), Enterprise Search tools and many others. Ultimately this will enable you to get more value out of your many technology investments, and drive a better, more efficient, effective and compliant business.

Knowing the data is the key to proving compliance. Part of information governance is showing what's going on in your data environment and proving any remedial actions taken to auditors.

First, NogaLogic gives you and your auditors a clear picture of what's happening in your data environment at any given time. Second, NogaLogic provides detailed reports which document the data environment before and after any remedial changes have been made. This can be critical if and when an incident occur, and can help protect you and company against charges of negligence. Third, NogaLogic's logs can also be used to prove any policy management actions taken on the files, such as data migration.

Knowing the Data is the Key to Governing it

Nogacom changes the way companies can and should go about governing their data – making it a far simpler and easier process than before. Nogacom's approach of classify first, plan earlier eliminates the significant challenges, complexity and fear that companies have when attempting to initiate an effective data governance strategy.

You no longer need to spend significant time and resources attempting to blindly guess the best way to go about it. Instead, with the insights gleaned through NogaLogic you now have a realistic picture of your data environment and the risks hidden within it, and can now start planning the most suitable strategies, the right resources and technologies needed to effectively and realistically address a wide variety of challenges such as regulatory compliance, PCI DSS compliance, Cloud Computing, eDiscovery, data migration and archiving, Electronic Data Records Management, and much more.

Data governance is not a one-deal, its a continuing and iterative process. With NogaLogic you can evaluate the ongoing effectiveness of your data governance policies and strategies and see to what extent users are complying with or working around them. You can also use NogaLogic to watch for unusual usage, trends or distribution of data. Then, if needed you can fine-tine your policies, adjust business practices, re-educate users, or make a conscious decision to accept certain risks in order to gain certain business benefits.

Your business is a constantly changing organism. So anytime there is a change – such as such as M&A activity, new regulatory mandates, infrastructure or technology updates, new business practices etc. - you should reassess your data environment and adapt your governance strategies accordingly, thereby ensuring that your company is always compliant, even as your organization changes and evolves.