BRINGING ORDER TO CHAOS: REGAINING CONTROL OF UNSTRUCTURED DATA

by Jacques Sauve

INTRODUCTION

Organizations have been accumulating data for years, sometimes decades. Corporate networks are storing literally millions of files, folders, and emails. As the demand for more storage grew over the years, there were two options to consider:

1. Clean up what was no longer needed and reclaim the space.
2. Add more storage capacity to the network systems.

Storage vendors reduced prices drastically, and we know that users are not very good at "cleaning up", and so option #2 was most often the solution. Humans are data hoarders; they will always say that they need to keep *everything*, "just in case". However, for organizations, this *should not* be a viable option. There are all sorts of issues to deal with around *information governance*, or IG – how to protect, optimize, and use the collective information as an asset.

Unfortunately, most organizations have lost control of their unstructured data repositories; over time, they have simply thrown more capacity at the problem without necessarily putting in place the policies, practices, or tools to bring order to the chaos.

The information that resides in that massive amount of data:

- ⇒ may have business value, but we don't know where it is, or that it even exists
- ⇒ may be requested in the case of litigation against the organization
- ⇒ may contain non-compliant information and put the organization at risk

External factors are putting more and more pressure on organizations to do something about the chaos: GDPR, PIPEDA, IIROC, and the Access to Information Act. Organizations are realizing that the status quo – keep everything and have no idea what's in there – cannot be maintained. There are key questions that need to be addressed:

1. Who can access our data?
2. What information do we save?
3. When should we delete it?
4. Where is my data being stored?
5. How do we audit everything?

An Information Governance strategy should address these answers, some through policies based on regulatory obligations (#s 2 & 3, for example), others through an assumption of knowledge: concerning #4, it is assumed that an organization's IT department would know *where* they are storing data.

The problem starts with tackling the mountain of electronic unstructured data; where and how do you begin? How can we determine who has access to our data, and what is in it (#s 1 & 5)? In this paper, we will present a simple approach that can help customers regain control of their unstructured data, which in turn will help them mitigate risk, benefit from cost savings, and put them in a much better stance when it comes to compliance.

THE VISION

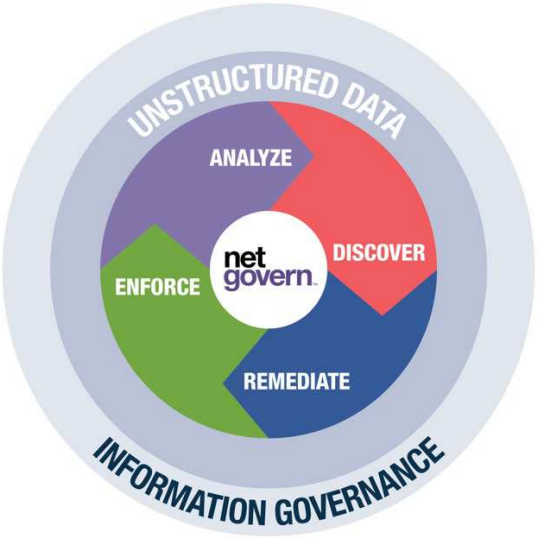
The ARMA vision of Information Governance, or Information Management, as they call it, includes a lot of moving parts. Essentially, it points out that Information Governance is a cross-functional discipline that involves many different stakeholders. It is a complex affair that requires much collaboration.

At the center of it all are the standards and principles that are used to guide the Information Governance initiative. This model provides a solid framework for developing an Information Governance strategy, as well as best practices.

At some point in their IG initiative, however, organizations still need a practical approach to dealing with the massive volume of electronically stored information, which we'll refer to here as unstructured data, as we've defined it above.

What should organizations do to regain control of their information? What's the logical process?

I am proposing an approach to tackling this issue that can be represented this way:



To be clear, dealing with unstructured data is only one aspect of an IG strategy, albeit troublesome and oft neglected one.

Within this sphere, customers need to focus on 4 different steps to regain control of this massive amount of information; we will examine each of these steps in the next sections.

ANALYZE

What types of files are stored in the file systems? How old are the files? How many duplicate files are out there? Who's hogging all the storage space? Which folders are the biggest?

From an access governance perspective, who has access to what files? What permissions do they have to any given location? How did they inherit those rights?

File analysis serves to identify the ROT – the Redundant, Obsolete, and Trivial data. Redundant can refer to the duplicate files scattered across the network or any instances of duplicate information. Obsolete data refers to files or information that is no longer valid; for example, old product brochures about products that no longer exist. Trivial files are those that have no business value whatsoever: music & video files, old Christmas party invitations, etc.

It is widely estimated that ROT can represent anywhere from 40-70% of the unstructured data in any organization; I've heard from some organizations that it's probably closer to 80% in their case! However, all this ROT is treated the same way as business-critical or sensitive information: it is protected, backed-up, and hosted on expensive storage.

Customers must also understand that any data stored on their network is subject to eDiscovery in the case of a litigation request. This means that all this ROT could potentially put the organization at risk! In the case of compliance & audit concerns, customers should be worried about what lurks in all that data.

Another aspect of what needs to be analyzed around all those files are the permissions. In many cases, over the years, IT has granted ad-hoc permissions to users and groups to fulfill collaboration requests. These permissions are rarely "cleaned up" and are very difficult to inventory. Customers need to gain clarity into these permissions when taking into consideration the security of their data.

The ultimate goals of the Analyze phase are to identify the ROT, gain clarity into the file systems, and understand all the permissions that have been assigned so that intelligent policy decisions can be made.

DISCOVER

Now that the organization has gained some clarity into the metadata surrounding their file systems, it is time for the Discover phase, which is where an organization would now look *inside* their data to deep dive into the content. This is where customers will discover the dark secrets that lurk in their unstructured data that can put them at risk.

There are many types of data to look for, and it will depend on the customer's industry. For example, in any organization that handles credit card payments, they would be striving to be PCI compliant. One of the constraints of PCI compliance is that credit card numbers should not be stored unencrypted in files or emails. How do we find those instances of non-compliance? How do we know that Johnny isn't storing a spreadsheet in his OneDrive that contains customer credit card numbers? Or that Joan hasn't been receiving credit card numbers through her email?

Customers concerned with external regulations (like GDPR, HIPAA, FINRA, etc.) would need to know if certain information is being stored on their networks. As previously mentioned, most have no way of assessing whether this is the case.

There is also the eDiscovery aspect of having control over unstructured data. If the organization is susceptible to litigation (e.g.: healthcare, construction, engineering, etc.) or Freedom of Access to Information requests, then it would greatly benefit them to be easily able to find that information, review it, and export it for the requesting party. Ideally, this would be handled internally by the customer's own legal department, rather than having to use expensive, outsourced firms.

REMEDIATE

By now a customer should have gained clarity into their unstructured data: their files, their emails, email attachments, network permissions, etc. They can now define Information Governance policies based on what they've found and what the regulations for their industry require. Based on their findings, their policies should cover retention schedules, access governance, and the defensible deletion of any data that no longer has business value.

Armed with this knowledge, they can take all the necessary actions to "clean house" and therefore mitigate any risk to the organization. Any ROT should be deleted, such as duplicate files, files that are no longer relevant to the business because they are so outdated, and non-business-related files (music, videos, old Christmas party invites, etc.). Old emails that go beyond the customer's retention schedule should be deleted as well, and emails belonging to users who are no longer with the organization but still within the retention period should be archived, where automated policies can take over to apply the necessary retention lifecycle.

This is where an organization would also decide – again, based on their policies – what to do with information that could pose a risk: those credit card numbers in emails and files, Patient Health Information that could be in those, as well. Do they delete them, or is there truly any business value to having them where they are?

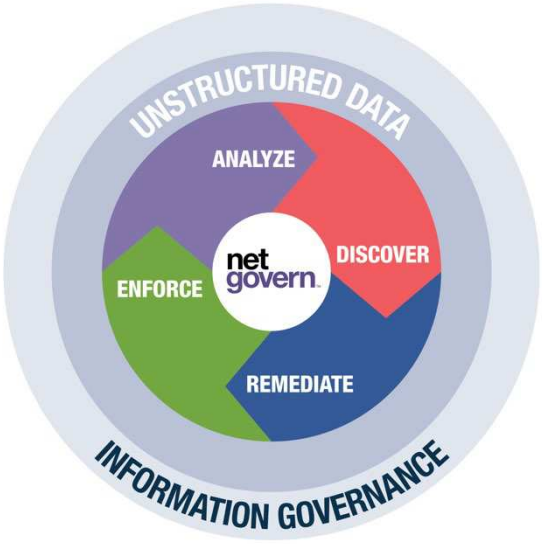
The organization should now also have a complete inventory of all the permissions assigned on the network, and there is a very high likelihood that a lot of remediation will be needed to reign those back in! This is a critical step in an Information Governance initiative, as it ensures proper protection of company data. Additionally, it can help prevent the spread of ransomware attacks, as infected workstations can normally only infect files that the user has access to; limiting users' rights strictly to what they need access to can go a long way towards mitigating the risk of having *all* data encrypted.

MONITOR

If an organization has gotten this far in the Information Governance cycle, they should – in theory! – now have much less data to worry about. It is not, however, the end of the road – Information Governance is not a "project" that is worked on and then checked off as "Done". Ever!

Certainly, the act of *starting* an Information Governance initiative, developing a strategy, and developing policies, *could* be seen as a project, but once in place, it is an ongoing effort. This can be likened to doing the spring cleaning in a home: if regular maintenance hasn't been taking place, it's a big job! Once completed and the house is impeccable, though, one is faced with a choice: let it get dirty, dusty, and filthy for a whole year, again, and do another big spring cleanup in a year, or do regular, weekly cleaning and always have a beautiful home! The latter also ensures that one is always ready for guests! It is also much easier to find things you need in a clean, well-organized home!

You'll recall that our diagram was in the shape of a *wheel*; that is entirely intentional and indicates that these 4 activities must be conducted on an ongoing basis. Akin, if you will, to the weekly house cleaning that one should do to make sure the home is always clean.



The amount of data versus the available resources in any given organization is disproportionate – that's how most of them got in the mess they're in in the first place! There's just too much data for humans to manage easily or properly.

Organizations should be using tools to automate most of these tasks and help make sense of it all. Delegating menial tasks through policy driven automation can keep the ROT and non-compliant data in check.

Setting up access governance policies and tracking those automatically against roles and responsibilities allows for significant reduction in risk. Monitoring sensitive information automatically will reduce risk and minimize the impact of security breaches. Alerts on GDPR, PCI, or other compliance information allows the security department to ensure that storage and communications policies for sensitive information are applied properly by the users.

Given the accumulation of data over the years – sometimes decades! – it is virtually impossible to undertake this task manually. Technology needs to be leveraged to help make sense of it all, clean it up, and then monitor the remaining data for any policy exceptions that could put the organization at risk. The status quo (keep everything and hope for the best) is simply not acceptable.

To begin with, organizations should be looking for solutions that could inventory all the metadata of the local file systems. This would allow them to gain clarity into "what's out there", who owns it, how old it is, whether there are any duplicates (and *where* all the duplicates are located!), what types of files are out there, who has access to what, etc. Based on those reports, organizations can then start formulating their policies around what to do with this data.

Customers also should be able to search *within* the data, so an eDiscovery or Audit solution that would index both their file systems and emails would be invaluable to discovering *what's in the data*.

Equipped with such tools, an organization should then be able to leverage these to remediate anything that is not compliant with their policies, and to automate the clean-up of the ROT.

And finally, technology should be able to help, on an ongoing basis, keep the unstructured data clutter-free and compliant with policies. For example, should permissions change on an area of the file system that contains sensitive data, custodians should be advised of the changes to ensure they are compliant with the organization's policies.

Regaining control of unstructured data often seems like a daunting task, but as Lao-Tzu said, "A thousand-mile journey begins with the first step."