Enterprise content management: securing your sensitive data

Simone Jeurissen
is Senior Consultant at KPMG Data & Analytics.
jeurissen.simone@kpmg.nl

Good enterprise content management is a must to secure your sensitive data. Especially given the astonishing pace at which the data volumes keep growing. This growth is driven by the digitization of society and accompanying new opportunities. The objective is to enable organizations to fully profit from (new) opportunities around data and be in control over the use of data. At the same time, laws and regulations (such as the GDPR in Europe, the CCPA in California and the FIPPA in Ontario) are becoming increasingly strict about what can and cannot be done with data. This is challenging for many organizations due to the messy character of large parts of the data. Gaining control over unstructured data is a tough challenge. Unstructured data has been built up for years and is often "hidden" within folders on file servers. How can organizations explore the potential of digitization and at the same time comply with data-related laws and regulations?

INTRODUCTION

"Governments of the Industrial World, you weary giants of flesh and steel, I come from Cyberspace, the new home of Mind. On behalf of the future, I ask you of the past to leave us alone. You are not welcome among us. You have no sovereignty where we gather."

These are famous lines from the 1996 declaration of independence by the libertarians, headed by John Perry Barlow. It was a time full of optimism about the societal effects of new internet technology and we had just started to explore the possibilities of this new cyberspace. The libertarians were predicting – or hoping – to build a new information Walhalla and even an independent republic, where governments and corporations had no influence. Almost 25 years later, we could not be further away from that scenario. The largest tech companies (FAANG, Facebook, Apple, Amazon, Netflix and Google) have changed our lives by changing the way we connect with each other and making information and products within hand's reach. In turn, they have gained massive power or even near monopolies, the web itself has turned very commercial and concerns about the proper use of personal data have become one of the major societal problems. Historically, technology has always had two faces. On the one hand, new technology offers opportunities for progress and innovation. On the other hand, new risks arise, such as the abuse of personal data. The challenge is to foster the positive, while controlling the darker side. The digital transformation is no exception to this. We have witnessed a multitude of useful and sometimes groundbreaking innovations that have made our lives easier and more comfortable. But it has also become clear that we must find (new) ways to deal with the dark side of digital transformation. Considering these factors, it is hardly surprising that governments stepped up their efforts trying to govern what in the early days was intended to be a sovereign place.

A DILEMMA

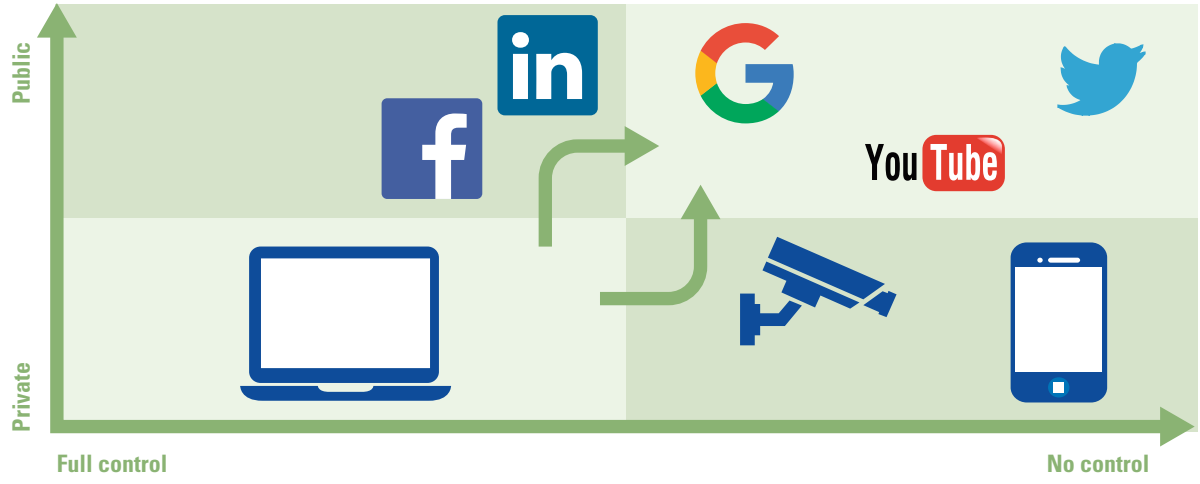
Some of the dynamics of this cyberspace are still valid. One of them: information wants to be free. Not only "free" as in "at no cost", but also "free" as in an "endless space to move around in". Professor Edo Roos Lindgreen once drew a simple graph with two axes to illustrate this ([Webw12]) (see Figure 1). One axis depicts the decrease of control over the accessibility of data; the other illustrates the extent to which data is publicly accessible. According to Roos Lindgreen, information follows the second law of thermodynamics: the result is maximum entropy. All information will end in the upper right quadrant of the graph. This is a situation of chaos: the information circulates freely in an uncontrolled space. If this model is valid, the evolution towards maximum entropy is inevitable, and data that arrives in the upper right quadrant can never be moved back. Take a viral video: once it is online it can never be fully taken off the internet. The saying that "you can't unscramble eggs" says it all.

Meanwhile, governments are trying to "unscramble eggs" in their efforts to regulate and improve the governance and security of personal data. In itself, this is understandable as we witness the risks and dangers of the unscrambled eggs nearly every day, such as the abuse of personal data. All in all, this leads to a dilemma where on the one hand information wants to be free and on the other hand laws and regulation aim to install boundaries to this freedom.

AT THE ORGANIZATIONAL LEVEL

The dilemma between freedom and control is challenging for society as a whole but is also valid for organizations. Many organizations are in the midst of a digital transformation. In this journey, they explore how they can

Figure 1. Accessibility versus control of personal data.



manage and profit from data. This often includes the need for freedom to innovate with data. However, stakeholders expect organizations to process data securely and be transparent about their data processing activities. Laws and regulations, such as GDPR, CPPA and FIPPA, limit freedom accordingly.

It may be tempting to opt for a quite liberal approach when using different organizational data sources, as this helps facilitate data-driven innovation. However, decentralized data processing activities require effective data governance. This governance is not only important to comply with laws and regulations, but also in order to warrant reliable and trustworthy data. Data governance ensures that all parties involved use one version of the truth to base business decisions on. The stakes are high: proper data governance is vital for success in a data-driven society, as being in control over your data means better information to facilitate decisions.

FROM DATA TO CONTENT

One of the solutions to deal with this dilemma is through master data management programs. Many organizations have created significant efficiency benefits and increased the level of information quality by implementing master data maintenance processes. This is because, thanks to these programs, master data objects are stored in *one* location. That way, other systems that make use of this information communicate with that one location, which serves as a single source of truth. Authorization management concerning these master data attributes only needs to be managed at the source, rather than in multiple locations. In these master data management programs, organizations clearly define which data objects are (strategically) important and implement structural management around these data objects. However, the challenge does not end there. The same principle should be applied to unstructured data as well.

As a result of digitization, data emerges from many new data sources. A significant part of organizational data is unstructured or semi-structured. Despite technological advancements that support the people who carry out business processes, the majority of these processes still require human interaction, and therefore the creation of content in some form. Examples of (semi)structured content are invoices, meeting minutes and photographs. Natural language is required to exchange information between business processes and parties. It is what makes these business processes human. Organizations must find ways to properly govern this part of the data pile too. Especially when it comes to personally identifiable and other sensitive information. That simply should not be dispersed over a chaotic unstructured data landscape.

We must find (new) ways to deal with the dark side of digital transformation

ENTERPRISE CONTENT MANAGEMENT

This is where enterprise content management comes in. Content refers to the data and information inside a container ([Earl17]). Examples of such containers are files, documents or websites. Content is of a flexible nature – it can change over time and knows its own lifecycle. Enterprise content management makes it easier to manage information by simplifying information retrieval, storage, security and other factors such as version control. The promise is that it brings more efficiency and better governance over information. Implementing enterprise content management successfully is not a walk in the park. The way content is structured is highly diverse – as it is often entirely dependent on the way of working of its author or the business process responsible. There are often few standards regarding the structure of information across business units. As a result, the majority of organizations still struggle with enterprise content management.

The stakes are high in a time of ubiquitous data where processing information has become a key differentiator: organizations with good information management practices make better decisions and thereby have an advantage over their competitors. This is not because they have all the information available, but because they are able to limit the amount of information to a relevant portion that human brains can deal with. American Author Nicholos Carr ([Carr20]) is one of many who argue that too much information might just destroy our decision-making capabilities.

Moreover, privacy and security (and the laws and regulations in these related domains) are two key drivers for better management of organizational content. The larger the volume of content stored, the greater the risk that it contains sensitive information, which could lead to reputational damage if it ends up in the wrong hands. For instance, shared folders often contain data extracts

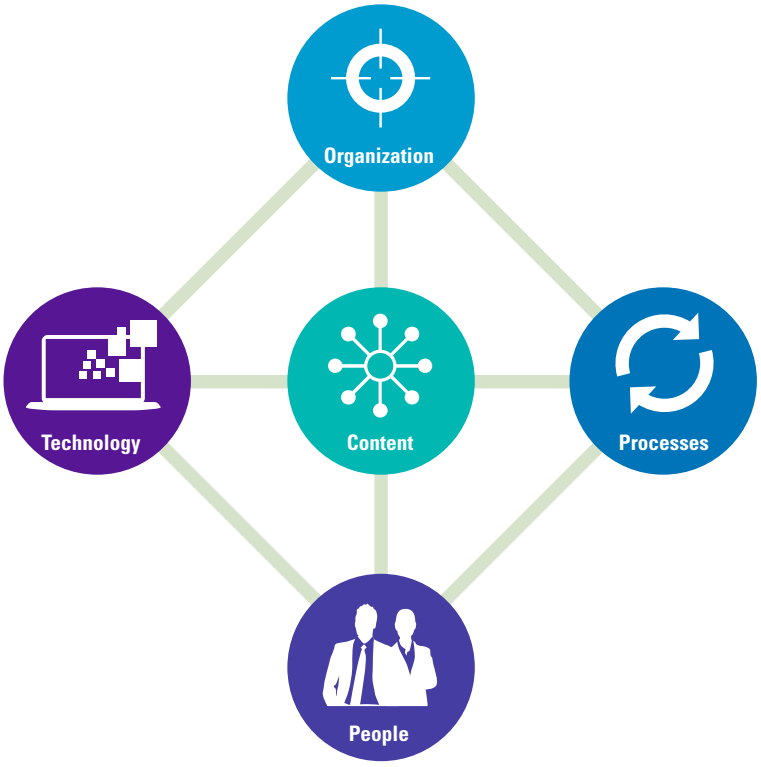


Figure 2. Enterprise Content Management model is comprised around five pillars.

from operational systems, which in turn, often contain personal data. What's more, a lack of enterprise content management also leads to inefficiencies. Traditional content management systems force artificial structure through folders. Without a strong search functionality, information retrieval is difficult when the exact storage location of specific content is unclear. Without proper data classifications, unstructured data is difficult to find, use, manage and turn into information.

ENTERPRISE CONTENT MANAGEMENT MODEL

Our view on content management is that it does not start with implementing tools and techniques. A holistic approach is used instead: a broad analysis of its relevance to an organization. To this end, we use an enterprise content management model based on five pillars: content, organization, people, processes and technology. This model is based on international market standards, such as DAMA DMBOK ([Earl17]), CMMI's DMM ([CMMI14]) and EDRM ([EDRM]), as well as the publications and experiences of experts in the enterprise content management domain ([Mart17]).

Content	
Components	**Description**
Scope	The content pillar determines the content that falls within the scope of the enterprise content management program. This requires the setup of an information or data catalogue (also referred to as a data dictionary), in which an inventory is kept of relevant content within scope of the program. For instance, customer contracts, product specification sheets and meeting minutes.
Controlled vocabulary	It identifies relevant content by defining a controlled vocabulary and indexing enterprise content for these terms. [Earl17] defines a controlled vocabulary as a defined list of explicitly allowed terms used to index, categorize, tag, sort, and retrieve content through browsing and searching. A controlled vocabulary is necessary to systematically organize documents, records and content. For sensitive personal data, a controlled vocabulary may exist of a list of illnesses. Controlled vocabularies are generally developed using taxonomies and/or ontologies that order terms. There is a strong link to information and data modelling, to define the objects and relations between them. There is also a strong link to the organization's business glossary: as the approved terms in a business glossary may be used in a controlled vocabulary as well. Other factors to consider are the use of synonyms and languages.
Records	Records are documents that need to be retained for a defined period of time once they have gained a formal status. Retention periods are often mandated by laws and regulations. For example, invoices that must be retained for 7 years.
Templates	Templates help standardize the structure of information across business units. Offering employees templates creates time efficiency gains, as employees do not need to create new materials from scratch. What's more, automatic metadata can be added to document based on the selected template.
Metadata model	The metadata model defines what information needs to be known about content throughout the organization. It refers to the properties of content, such as the authors, creation data and status.
Media types inventory	The media type inventory defined the accepted media types within an organization. This is closely linked to the approved applications within an organization. For instance, for an organization that only uses Windows, no Mac-related files would be expected.
Authorizations	Authorizations determine who has access to what content, based on the content rather than the application the content is located in. This way, it is ensured that only the right people have access to the right information.

Organization	
Components	**Description**
Policy	The data and/or content management policy prescribes the guidelines and rules (e.g. regarding privacy and security) that apply to the handling of data and content throughout the organization. For example, that all transfers of personal data must take place through approved channels in an encrypted form.
Governance structure	The definition of roles and responsibilities, ownership structures and the commitment of stakeholders. The goal here is to govern data, content and information as a strategic asset. This includes the boards and approval structures needed to validate content. These are often documented and formalized in the form of RA(S)CI's.
Funding	A funding model must be defined in order to provide financing for development of organization-wide content management standards; such as new templates, required metadata fields and storage environments. This is a must to meet changing business requirements.

Processes	
Components	**Description**
Business requirements	Business requirements define the wishes and demands of business users. These should be clearly defined for content: what information is required and when, who is allowed to see what type of data? In what format? Business requirements not only arise from business users, but also from laws and regulations (legal requirements) and security standards (security requirements).
Process mapping	Content is related to processes, which is why processes that consume content should be mapped. This allows for the identification of possible content creation workflows and approval moments. Process mapping helps identify business requirements. Factors to consider here are process KPIs to improve content handling activities, as well as process controls to ensure these activities meet business, legal and security requirements.
Lifecycle	The lifecycle of content defines how content evolves over time, like through versioning: when does a document become a 1.0 version? Or archiving: when is a document moved from its active location? It is recommended to use standardized document statuses throughout the organization, examples are: draft, formalized and obsolete. This helps indicate the quality of content.

Technology	
Components	**Description**
Architecture	The architecture required to support people, organization and processes must be defined. Key elements to consider here are integration between content storage environment (to avoid duplication for example) and backup and recovery. Depending on the value of content, stricter recovery possibilities may be required. In architecture, privacy by design can be applied as a principle. For instance, by centralizing personal data and applying pseudonymization to personal data when it is used in systems that do not necessarily require this information to function. The chosen architecture should be defined using business requirements (based on the processes pillar).
Metadata tagging	Tooling should support data classification and the tagging of content, as prescribed by the metadata model (from the content pillar). For example, classifying content containing financial information that is "sensitive", so that appropriate mitigating actions (such as restricting access) can be taken. Metadata tagging refers to the technological functions required to fill out the metadata values, such as the media type of a file. Key is to make these metadata fields fully searchable.
OCR	In order to manage flat files, Optical Character Recognition (OCR) is required. Using OCR, flat files can be "read" to make these fully searchable on text. Think of a scanned contract, for example.
Monitoring	It must be possible to monitor the content landscape on higher levels, to identify trends, such as non-compliance to content management policies, but also for innovative purposes. Such as topic modelling and content curation, to stimulate knowledge management.
Search	Effective content management requires tooling, to index organizational data sources, including shared network folders, such as SharePoint sites and e-mail archives. Indexing tooling makes data sources fully searchable, allowing for the quick identification of personal data. Some indexing tools also have content production functionalities, which can be used to automatically redact personal data and publish the information in a format so that the personal data is no longer retractable. If needed, an original copy may be stored in a secure location. There are many indexing tools available on the market, [Eijk18] wrote an article illustrating the functionalities of several alternatives.

People	
Components	**Description**
Culture	People play a large role in the development of quality information, the organization should provide the means to support this. Culture is a complex concept to grasp, [KPMG16] created a model to apply soft controls within an organization. For example, by identifying which interventions are necessary to create a content management community. Publishing employee success stories for example (role modelling). In an effective content management community, employees work in a consistent manner and actively contribute to the effective management of information and data, for instance by initiating data-driven projects. At the end of the day, people have the greatest impact on how content is managed.
Training	In an ever-changing technological landscape, it is important that employees develop the skills required to effectively manage content. For example, by training employees in the use of the organization's document management system or eDiscovery tooling.

The model is valid for any data architecture. Even in an extremely traditional organization – with content stored in paper files – the approach will trigger the right questions and will lead to a well thought-out solution. Information that is written in natural language can be digitized, as is also the case with photos of letters. OCR enables interpreting and managing the information on these flat documents. The same goes for images: artificial intelligence helps understand images. In fact, over the years we have all contributed to that by validating that we are not robots. Millions of users that perform the Captcha test have trained these algorithms ([OMa18]).

GUIDELINES FOR SOLVING THE DILEMMA

As described earlier, we have a dilemma at hand: there is tension between freedom of information and the need to govern all this information. The aforementioned model helps to define a holistic approach, and by exploring the five pillars, organizations enable a tailor-made approach that suits their specific characteristics and challenges.

There is tension between the freedom of information and the need to govern all information

The following general guidelines may be helpful in dealing with the dilemma.

1 Create awareness that content management is more than compliance

In practice, many organizations explore the options of content management in order to deal with privacy and security concerns, triggered by laws and regulations. This is understandable as the stakes are high and media attention for non-compliance issues exists. However, a better recipe is to start with the virtues of content management for the organization. In the current business landscape, being data-driven is key to success. This means that there are great benefits in developing a controlled vocabulary and curate con-

tent about specific topics. Content curation around specific topics that are important to an organization can stimulate innovation and knowledge management. Once the value of this is recognized, it will be easier to keep up in terms of compliance.

2 Use the full potential of clever indexing tools
Indexing tooling offers great opportunities to index content that is stored in a variety of systems, even the highly unstructured ones such as shared file shares, SharePoints and OneDrives. Especially when opting for a decentralized approach, indexing tooling offers quick methods to identify personal information throughout the organization. For example, by using a regular expression for bank account numbers, they may be quickly identified. A regular expression is a sequence of characters (e.g. numbers or letters) that defines a search pattern. To illustrate, to find all Dutch telephone numbers, one could search for strings of 10 numbers long that start with "06" or "+31". This technique is developed in theoretical computer science and formal language theory. Possibilities go much further than that, however. The application of entity extraction for example, allows for the quick identification of people, places and concepts that may be deemed sensitive. Executing the "right to erasure" is difficult to implement when business processes involve high volumes of content, such as customer letters, data extracts in Excel format and emails. Implementing content maintenance processes mitigates this problem. All personal data can be identified quickly, and the correct follow up action can be taken. Mature organizations automate these processes, based on set data retention periods.

3 Opting for privacy by design
A centralized approach for storing personal information offers better options for the governance of information. Personal information should not be scattered across network servers. By centrally storing personal information, risks are reduced as there is only one location for the application of governance rules. In other systems, pseudonymization techniques can be used to mask personal information. In practice, many of these applications have no need for information that can be traced back to an individual. This central approach creates flexibility to use information in decentralized applications. This way, the group of employees who do have access to personal information can be limited to the employees who truly require access; the customer service department for example. For other activities, such as the management of transactions of the analysis of customer behavior, personal information is removed or pseudonymized, making it impossible for the users to trace that information back to an individual.

By centrally storing personal information, risks are reduced

Case: GDPR triggers data retention program in bank

As a result of GDPR, a bank had made steps regarding data retention, heavily relying on their employees to cleanse sensitive data. New policies were created to determine what data was collected and for how long it was retained. One of the issues concerned the fact that content can contain multiple types of personal data. Take a CV for example, it contains a name, telephone number, address and sometimes even a date of birth and/or a picture of a person. Because content, such as CVs, were stored on shared file servers and within email boxes, it was difficult for the data privacy officer to quantify the success of the steps they had taken.

We helped quantify efforts: we carried out an analysis to determine how much personal data was left. We analyzed a total of approximately 100,000 emails and 1,000,000 files. 60% of the content we found was redundant, obsolete and trivial (ROT). ROT content is content that no longer needs to be stored, as it does not have business value. A common example: duplicate files, e.g. multiple copies of the same manual stored in different locations. The oldest file in the analyzed dataset dated back to 1997. We found meeting notes from 17 years ago containing client information. Even employee notes calling a customer "very sweet" and another customer "very annoying". The list goes on, we identified 6,000 social security numbers, 200 CVs and even 500 files containing personal medical information. We created lists with files to be cleansed, these were validated by the business and then automatically deleted by the IT department using an automated script. Within 2 weeks of work, we had reduced remaining personal data by 50% and identified next steps to get that number down to 0%.

Case: Professional services migrates to the cloud

A popular topic is phasing out file shares and moving to the cloud. A professional services company had this same ambition. Their question, however, was how to approach this migration to the cloud. They faced several challenges. Authorizations management on file shares was not effective, due to the use of many different user groups over time. Different user groups as well as individual users had obtained access to specific shares and folders, making it very difficult to determine the owner of specific content. As a result, it was not possible to ask the right owners what data should or should not be migrated to the cloud. What's more, these authorizations could not be copied to the new environment, as they were no longer up to date. An entirely new authorization concept and structure was required. We helped this client carry out their migration by utilizing technology to simplify the process. We classified existing data, around cases that made sense to the organization's operations. In this case, these "cases" were projects, clients and departments. For each project, client and department, new environments were created, and the relevant files were migrated to that environment. Files with sensitive information were automatically classified using regular expressions for personal data. The information within a file, was automatically recognized and redacted upon migration. A new version of the document was created, in which the sensitive information was blacked out. It was no longer readable nor retrievable by the end user. The original stored in a secure location, for a predefined period of time to make sure no valuable information would be lost.

CONCLUSION

In the current era of ubiquitous data, organizations face a new dilemma. On the one hand, information wants to be free to explore the (new) opportunities of this era. On the other hand, the (messy) information within an organization needs to be controlled. Laws and regulations have raised the bar in recent years. It is a complex challenge. The good news is that there are a number of promising techniques and concepts that help organizations deal with this complexity. Organizations that start with defining the benefits of content management – having a controlled vocabulary, better insights for decisions, improved knowledge management – are best prepared to deal with this dilemma. Our model offers them a guiding hand.

References

[Carr20] Carr, N. (2020). *The Shallows: What the Internet Is Doing to Our Brains*. New York: W. W. Norton.

[CMMI14] CMMI Institute (2014). *Data Management Maturity (Dmm) Model* (1.0 ed.).

[Earl17] Earley, S. & Henderson, D. (2017). *DAMA-DMBOK: Data Management Body of Knowledge*. Bradley Beach, NJ: Technics Publications.

[EDRM] EDRM Model. (n.d.). Retrieved on December 15, 2019, from: https://www.edrm.net/resources/frameworks-and-standards/edrm-model

[Eijk18] Eijken, T.A., Molenaar, C., Dashorst, I.M., & Özer, P. (2018). eDiscovery Spierballentest. *Compact* 2018/2. Retrieved from: https://www.compact.nl/articles/ediscovery-spierballentest/

[KPMG16] KPMG (2016, February). *Acht basis soft controls.* Retrieved on January 1, 2020, from: https://assets.kpmg/content/dam/kpmg/pdf/2016/04/20160218-acht-basis-soft-controls.pdf

[OMal18] O'Malley, J. (2018, January 12). Captcha if you can: how you've been training AI for years without realising it. Retrieved on December 12, 2019, from https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it

[Mart17] Martijn, N.L. & Tegelaar, J.A.C. (2017). It's nothing personal, or is it? *Compact* 2017/1. Retrieved from https://www.compact.nl/articles/its-nothing-personal-or-is-it/

[Webw12] Webwereld Redactie (2012, March 26). Eén grote vrijwillige privacyschending (opinie). Retrieved on December 12, 2019, from: http://webwereld.nl/social-media/59974-een-grote-vrijwillige-privacyschending-opinie

About the author

Simone Jeurissen is Senior Consultant at KPMG in the Data & Analytics unit. She specializes in Enterprise Data Management. She advises her clients on the design and implementation of the technical and organizational mechanisms required to produce and maintain high quality (un)structured data sources.