Why not let the data do the talking?

Simone Jeurissen MSc
is a senior consultant at KPMG Data & Analytics.
jeurissen.simone@kpmg.nl

Pieter Scherpenhuijsen
is CTO of Indica.
pcs@indica.nl

Organizations have made great efforts to make their processes faster and leaner by 'going digital'. Faxes have been replaced by emails, dossiers have moved from desk drawers to the cloud and paper forms have turned into iPads. Digitization has exponentially increased the volumes of unstructured data. The findability and accessibility of data is dependent on the metadata added at the source. Standardizing metadata controls is time-consuming and costly as it is difficult to tailor to the needs that arise from differences between processes, systems or even departments. Unstructured data ownership is often undefined ([Mart17]) and in turn, data is continuously created without defining who is responsible for its deletion. These issues lead to the creation of an unmanageable ocean of unstructured data. A smart data platform overcomes these issues by using text analytics to create automatic metadata-driven context around unstructured data sources.

On average, more than 56% of digital libraries consist of redundant data

Table 1. Terms & Definitions.

Structured data	Structured data refers to data that has a defined structure (data model). These are tables that fit within relational databases.
Unstructured data	Unstructured data refers to any data that does not have a recognizable structure. These are often files that contain text or multimedia content. Examples are documents, emails and images.
Text analytics	Text analytics refers to the set of linguistic, statistical and machine learning techniques that model and structure textual data sources.
Digital library	A digital library refers to all available data sources within an organization's IT landscape.
Data retrieval	Data retrieval refers to the way business users can find, extract and use data from systems within the IT landscape.
Data retention	Data retention defines the policies for meeting legal and business data archiving and deletion requirements.
Data control	Data control refers to extent to which an organization has insight into the content of the data that they process and where this data resides.
Redundant data	Redundant data refers to data that is no longer needed for legal or business data archiving requirements.
Data archiving	Data archiving refers to the process of moving data that is no longer used daily to a separate storage location.
Metadata	Metadata refers to information about specific data, for example author, date modified or classification of data.

INTRODUCTION

Unstructured data plays a crucial role in business processes. When creating reports and overviews, manual data extracts are made and stored somewhere on a shared file server. When carrying out an inspection, a photo is made for evidence. When preparing for a meeting, emails are exchanged to create an agenda. Excel sheets, photos and emails are just a few examples of unstructured data. Over recent years, organizations have made great strides in digitizing processes, through systems, applications and even robotics. Digitization has led to the creation of even more data, moving away from traditional paper and electronic documents alone. Digital libraries now also include open text fields, emails, chats, videos, images, scripts, tweets and many more. When not effectively managed, these data types often hinder process efficiency objectives.

This article dives into the way a smart data platform employs text analytics to automate metadata creation for unstructured data, eliminating the need for controls at the source. It starts out by explaining the challenges that arise from unstructured data. It then goes on to explain how managing an organization's digital library is different when there is no librarian, as opposed to a situation where the smart data platform takes on the role of the librarian. This is followed by a detailing of the benefits of a smart data platform implementation, namely: better data retrieval, retention and control. Then, the text analytics technologies used for metadata generation are discussed in more detail. Finally, the steps needed for the implementation and maintenance of a smart data platform are explained.

UNSTRUCTURED DATA AND ITS CHALLENGES

Unstructured data is hard to manage, because it is difficult to define what it contains, especially when dealing with large volumes. Without understanding what it contains, its relevance and value for business processes cannot be defined. Three main challenge arise from a lack of data management around unstructured data sources: ineffective data...

1. retrieval;
2. retention;
3. control.

Data retrieval is challenging, because business users must often access data in inefficient ways. Take folder structures for example. Although folder structures can greatly help users to create structure for themselves, they are often indecipherable to anyone other than their creator. This is because it is hard for other users to understand the logic behind folder structures when only looking at folder names. Viewing what is stored inside folders takes up a lot of time. When data is found, a copy is often stored elsewhere to improve future findability. This behavior increases unnecessary storage costs even further. Even in scenarios where retention-worthy data has a designated space in an application, additional copies are stored elsewhere 'just in case'. In turn, it is unclear which version is the last. The larger the size of the digital library, the more time employees spend looking for specific data they need to carry out business process activities.

Data retention and deletion is necessary from both a business and legal point of view. When data needs to be retained for legal purposes, and is no longer needed for daily operations, it should be moved to an archive. When data is no longer needed for daily business operations, and does not need be retained from a legal point of view, it should be deleted. This is often not the case, creating large amounts of redundant data in the digital library. Retention and deletion is challenging because of a lack of data ownership. Since no one is made responsible for deletion, data is stored longer than required. A folder for a specific contract could contain a hundred draft versions of that same contract. Retaining the folder may seem logical, but there is only one version of that contract that has true value. A lack of retention and deletion leads to the situation where the digital library becomes so large that deletion is perceived as impossible. This is due to the fear of accidently deleting something valuable. When dealing with large volumes, it is hard to link retention periods to specific data. Storing data without applying data retention rules leads to a failure to comply with laws and regulations around data privacy and data security. This is because it is challenging or even impossible to determine whether data retention periods have passed. Experience in the field shows that on average, more than 56% of digital libraries consist of redundant data.

The final challenge is data control. Since data is stored in numerous decentralized locations, there is no clear overview of what is available within the digital library, and where that data resides. As a result of continuously increasing volumes, data control becomes more challenging every day. Even when controls do exist, they cannot be carried out without knowing where data resides. Think of GDPR compliance for example: financial reports or customer service letters may contain personally identifiable information. This data must be deleted if its retention period has passed. Otherwise, authorizations to access the data need to be restricted to a need-to-know basis. However, due to the lack of insight, action cannot be taken. What's more, it is often unclear who has access to what data. There is not enough tangible evidence to show data owner's efforts with regards to data management. In turn, it is simply unclear whether departments do or do not control their data effectively.

All in all, these issues suggest the need for better management of unstructured data sources.

A digital library without a librarian

Imagine an organization as a digital library, where business users have taken the role of the authors and readers, but there is no librarian. All books are placed in bookshelves that define their genre, whether it is science fiction, history or romance. The bookshelves create structure, to make books findable. However, this does not work the way it was intended to. When an author writes a book about a historic hero, they define it as 'adventure', and create an adventure bookshelf for it. The author does not know that perhaps another author has already created an adventure bookshelf elsewhere. On top of that, the content of the author's book is ambiguous. In fact, the book could be placed on two shelves and not just one. When looking for that new book about this historic hero, it is unclear whether the reader should look for 'history' or 'adventure'. Sometimes, a reader is looking for something for which no bookshelf exists at all, such as a specific writing style or a type of imagery. All in all, readers are often left confused as where to look and find the books they are interested in. They ask fellow readers for help and eventually find what they are looking for, but they lose a lot of time in the process.

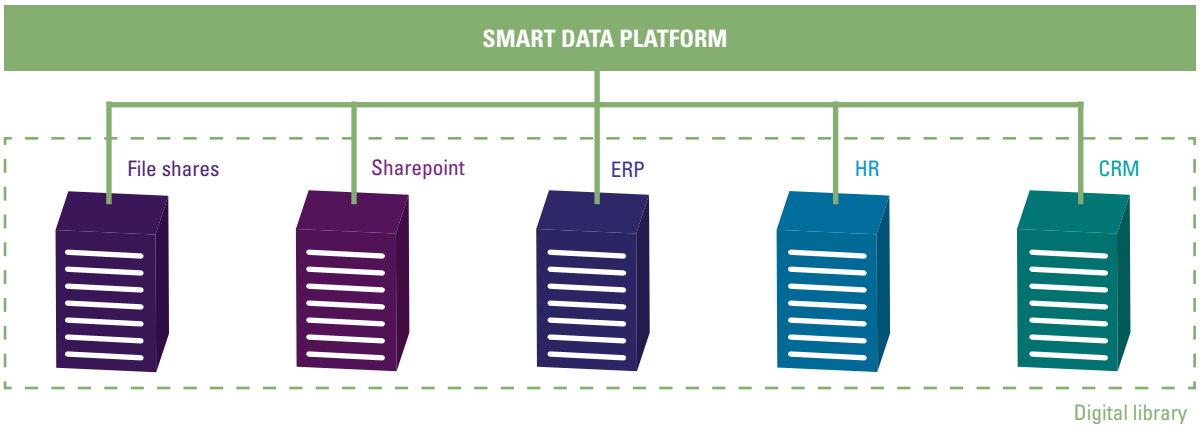

Figure 1. How the smart data platform indexes the digital library.

The Smart Data Platform as the Librarian

Now imagine a digital library where the smart data platform takes up the role of the librarian. Instead of making authors place their own books on shelves, this is done automatically. The librarian takes the information provided by the authors, as well as the books' actual contents to determine on which shelves the new book fits. This is done only after comparing the new book with all other books readily available in the library. Books are labeled based on their similarities, for example their writer, genre, language or even their writing style. What's more, the librarian makes use of virtual bookshelves. So instead of making readers find books on one predefined set of shelves, the librarian can arrange the books in different sets of shelves. The librarian will always present books in a way that makes the most sense in the context relevant for the reader. That is, if the reader asks for genres, the books are arranged by genre. If the reader asks for languages, the books will be arranged by language.

> With the smart data platform as the librarian, readers always find the books they are looking for

In this metaphor, the books are unstructured data and the bookshelves are virtual buckets defined by metadata. Note that these are *virtual* buckets; data is not restricted to a single bucket. These buckets are defined by a combination of metadata fields. Different combinations of metadata fields make data presentable in numerous ways. With the smart data platform as the librarian, readers always find the books they are looking for.

Improving Data Retrieval

The smart data platform improves data retrieval in two ways: by making data fully searchable on content, and by making data accessible through numerous buckets. Data can also be accessed without using buckets, but purely by searching for specific content, because data is not bound to any bucket. A user's authorizations on the platform are the same as in source systems. The platform reads access rights from connected systems and mirrors these authorizations in the platform. Since data is fully indexed, a user could remember one specific sentence from a fifty-page contract, type this into the search engine, and find it. Users can also search and find content by entering key words that make sense to them. If the user chooses to search for that contract by using the supplier name, they will then be presented with all data that contain information about that supplier. Instead of storing twenty copies of the same data in different locations, the same copy is stored in one location, and made accessible in twenty different ways. When searching for that specific contract, not only the specific Word document is shown as a result, but also the PDF version and all email communication related to that contract negotiation.

Improving Data Retention

The smart data platform facilitates data retention, as generated metadata can be used to make informed decisions about what to retain and delete. Data that needs to be retained from a legal point of view can be identified by the data class. There are specific laws about the retention of

invoices, personnel dossiers and audit reports, for example. How long, from the moment of creation or formalization, should it be stored? For data classes with a formal status, retention is automated by having a periodic automatic deletion of all data that has exceeded its set retention period. For data that does not need to be retained for legal reasons, the platform helps with retention as well, by determining when it becomes subject to deletion. Data can be deleted not only based on its age, but also based on its content and/or how often it is viewed. It is much easier to say 'yes' to the question 'may this be deleted?', when you understand the content and how often this data is accessed by business users. The platform helps organizations automate retention rules through metadata.

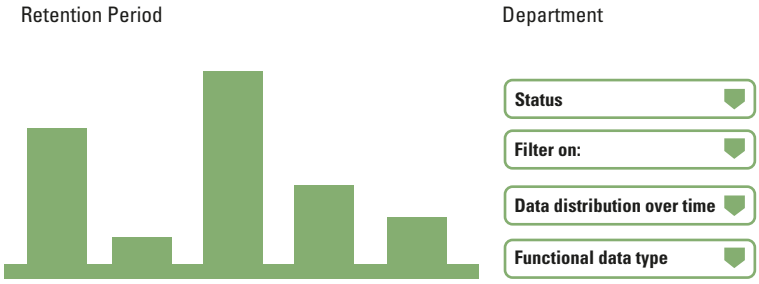

Figure 2. Example of metadata-driven insights into the digital library.

Improving Control

The smart data platform helps regain control over the digital library, serving as a data management platform. Data control moves from being a purely IT storage cost-driven task to a business matter. Using the platform for data control provides four main benefits:

1. The platform provides an overview of the entire digital library, as it can be connected to all digital sources. The platform provides technical and content insights into these sources through metadata. The platform contains modular dashboards that allow data owners to filter on these overviews, allowing them to answer specific questions. Think of checking how many new contracts were created in the past year, for example.
2. That it offers a method to automatically filter out unlawful data. Unlawful data could be a national security number in a marketing folder, or a copy of a customer's passport in a public folder. Intelligent search queries and text analytics can be used to filter out this specific data. Not only does the platform provide a general overview of the digital library, it can also be used to ask extremely detailed questions. Think of finding financial overviews that contain bank account numbers, for example. To help with GDPR compliance, data privacy officers can be given authorizations to monitor the use of personally identifiable information. Unlawful data that should no longer be stored as mandated by policy can be marked for archiving or deletion.
3. That it gives clear insight into authorizations. A platform user can view all data they are allowed to see, without interfering with the source systems the data is extracted from. Authorizations are managed on a data level, instead of on a folder or application level, making specific content available to individuals who require access. Access rights are given in source systems, but can be monitored through the platform. This helps detect sensitive data sources too many users have access to, as well as users who have access to too many data sources. As multiple source systems can be connected to the platform, the platform gives insights into the entire organizational authorization structure.
4. The platform makes data ownership tangible. Data owners are given read rights to data stored on their file shares and systems. They are made responsible for all data within their department ,and the platform offers clear insights into their data. Managers can report on the exact amounts of risks in terms of personally identifiable information, but also quantify business process output, in terms of the amount of customer letters or contracts created per period. The platform gives management something tangible to quantify their efforts, and KPIs can be based on these insights.

It is much easier to say 'yes' to the question 'may this be deleted?' when you know what the content is and how often this data is accessed by business users

Practical examples of how unstructured data influences business process efficiency

Without the smart data platform
Unstructured data created within the R&D department is the starting point for other organizational processes, including the sales department. The sales department needs access to R&D data, as it is required to create sales material and target customers. However, R&D and Sales are autonomous departments. They each use different systems between which no communication exists at all. So, when a new product is created, the parameters and details of the product are inserted in an Excel sheet and sent to the Sales department by email. The Sales department employs one person who manually enters the data from the Excel sheet into the Sales system. Sales often requires additional data for their processes, which is not entered into the Sales system, as it is not part of the Excel template. Sales representatives individually call their contacts within the R&D department to ask for additional data. R&D employees spend 25% of their time answering phone calls.

With the smart data platform
The Sales department searches for new product data in the smart data platform. They find all information that is known about the product. Search results are sorted by date, so they make use of the latest information. The R&D department is no longer hindered by the many phone calls from Sales. The employee in the Sales department who used to enter data manually, can use this time for more value adding tasks.

Without the smart data platform
Actuarial processes within large financial organizations are an unstructured data hazard. In order to make essential calculations needed for risk assessments and such, dozens of Excel spreadsheets travel through the organization by email. The reason these spreadsheets travel by email is because employees work from different storage locations their colleagues either do not have access to or do not understand due to their complex folder structures. Tracking the flows of data through these spreadsheets, many duplicate values and circular references are detected. To the point that if these numbers were to be integrated in one Excel sheet, most of the tabs would be redundant.

With the smart data platform
Spreadsheets no longer need to be sent by email, nor do employees lose time trying to understand complex folder structures, to determine where spreadsheets are stored. Spreadsheets are saved in one location, and access rights are given to all actuarial employees who need the spreadsheets for their work. The latest version is always shown at the top of search results. Other, similar calculations are also suggested as search results, to quickly identify duplication.

Without the smart data platform
Sometimes 'going digital' does not work out the way it was intended to. The system that employees need to use for their business activities does not work the way they would expect it to, or simply does not meet their requirements. An example is a company that has a limited amount of assets. Yet their system cannot generate one full overview of all these assets. Over the years, employees within the finance and control department have created a list that does provide this one overview. However, this list has grown immensely over the years, leading to an Excel file of 50 GB and a hundred tabs. It gets more serious than that even. Since the file is so large, employees cannot share it with each other through their business email. When they do want to share it, they turn to the use of their private email, which allows larger files to be sent across servers.

With the smart data platform
Instead of trying to create a full list of assets in a system that does not offer that functionality, a list is created and maintained in the smart platform. All data related to an asset is bundled around that asset, recognized by text analytics around the asset name and numeric identifier. The platform recognizes when a new asset is created, and automatically adds it to the list. The risk of a data leak is reduced, as the Excel file no longer needs to be sent by mail.

METADATA GENERATION IN THE SMART DATA PLATFORM

A smart data platform takes a different approach to metadata creation. Instead of entering it manually at the source, metadata is generated automatically using text analytics. The smart platform consists of an indexing functionality, which reads all connected data sources, and makes a mathematical representation of their content in the index. This mathematical representation determines what the unique characteristics of a data source are (e.g. the frequent and non-frequent terms), to map data in comparison to other data sources within the digital library. The indexer looks beyond applications and folders by extracting text from its source. Next to the indexation of content, the platform also collects all other technical metadata from source systems. For documents this could be the date last modified, the extension type and the file path. For images taken by a smartphone, this goes as far as to include the geolocation of the photo. The four main technologies that are used by the platform are regular expressions, pattern recognition, rule-based searches and classifications. These are explained in the following section.

A regular expression is a sequence of characters that defines a search pattern. Regular expressions look for specific text strings. The mention of a specific text strings goes to great lengths in identifying valuable information. Think of searching for bank account numbers, telephone numbers, or product codes. If found, this information is automatically extracted and added as metadata.

Pattern recognition not only looks at the text, but also the spaces within data. This way, it can recognize data that complies with a standard format. This is often the case for records, such as contracts, letters and invoices. Records are data that must be retained, as mandated by policy, laws or regulations. Pattern recognition identifies data classes and adds this information as metadata.

Rule-based searches look for the presence of specific information, often in the form of a list of (master data) values. This list can either be uploaded once, or be generated by a connected structured system (e.g. CRM, HR, ERP). Master data from a CRM system can be employed to identify all data relating to a specific client or supplier. Master data from an HR system can be employed to identify all data relating to an employee or rejected applicant. Data is scanned on whether it contains one or more of these values. Depending on the result of the search, a specific metadata value is added in the platform.

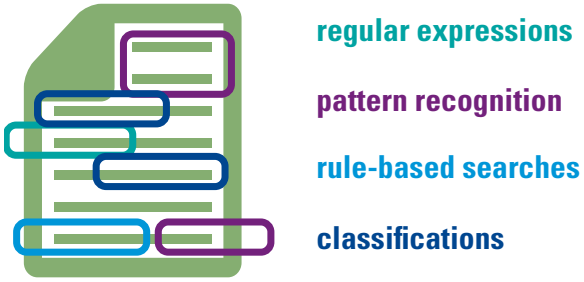

Figure 3. The text mining technologies used to identify specific data.

Classifications group similar data. They are made based on a training set. This is a small set of the data that needs to be found within the digital library. Think of resumes, for example. The system is given a small group of resumes, and uses this training set to recognize the rest of the resumes. Classification is a key technology to automatically identify valuable data within large digital libraries.

Based on results of the analysis, data is automatically labeled with metadata, such as a data class; an audit report, customer letter or python script. Thanks to the metadata, users will no longer need to open and fully read data to understand its value and relevance.

IMPLEMENTING A SMART DATA PLATFORM

A smart data platform implementation requires an enterprise data management implementation program that will take up several months, depending on the size of the digital library. The implementation program consists of two parts, a technical and an organizational program. The organizational program requires the creation of a data management policy, governance, data lifecycle processes, templates and controls, where these are not available yet. Governance includes the assignment of data ownership. The technical program consists of the installation and configuration of the platform.

The first step in the technical program is to use insights provided by the platform to get rid of redundant data. The platform uses text analytics to scan through all the data and create collections of redundant and valuable data.

Redundant data is classified by filtering out data that is no longer relevant from both a technical and business point of view. Examples of redundant data from a technical point of view are duplicate data, empty data or corrupt data. Examples of redundant from a business point of view are data from customers who have not been a

customer for longer than seven years, data about an application that has been phased out, or draft versions of a document that has been formalized. Redundant data makes up most of the digital library, leaving a relatively small amount of valuable data after the initial clean up.

Next, valuable data is classified. This is done using the four technologies described earlier, in combination with business knowledge from the organization. That is, the platform determines which data appears frequently, and business users are asked to give that data a functional name. The result is a cleansed digital library that only contains relevant data, that complies with a minimum set of metadata requirements.

The platform contains metadata generation rules. The maintenance of these rules is responsibility assigned to data owners within the business. When business processes change, the data owner creates new (sets of) metadata rules to make sure new data has a place in the platform. The platform will generate suggestions for new metadata; these suggestions are validated by data owners. All metadata rules are centrally managed by data owners. Business users can also enrich existing metadata with their own search terms. The platform will use newly added user information to create new metadata rules, continuously offering new and improved ways to view and find data. In turn, the platform serves as a long-term mechanism that offers one central location, that users can use to find anything they may be looking for. Organizational choices regarding systems and tooling for business processes are independent of the functioning of the platform, as only the extraction of text is important, leaving the organization free to innovate with technological advancements. When the platform finds data that does not fit within predefined metadata, it will signal this to the relevant data owner. They are then able to create new metadata rules, or help the platform understand where the data belongs. The platform will learn from this information, and apply it to all future data that is added to the digital library.

CONCLUSION

Digital organizations use data to create new value and insights. The implementation of a smart data platform can greatly help with the digitization of an organization, as it extracts data from source systems and maps it in the digital library. Data management through a smart data platform creates great benefits for organizations; improved data retrieval, retention and control. Its implementation requires little effort, as it brings no major changes in the way of working employees have grown used to. The platform helps with the implementation of data ownership, by making efforts tangible through quantitative insights and reports. Further development of business logic through virtual bucket creation is carried out by users and data owners within the business, making data management a business-focused task. All new information fed to the platform is further applied and standardized through the platform's text analytics capabilities. Many possibilities lie within the further development and automation of smart data platforms. An interesting application in the future would be to deploy such a platform on file servers containing data analytics scripts. Data scientists could search for specific algorithms or functions, to find other applications that could benefit the development of their own analyses. The modular design of the platform and such integration mechanisms ensures that organizations can adapt to any changing needs in the future.

Reference

[Mart17] N.L. Martijn and J.A.C. Tegelaar, *It's nothing personal, or is it?*, Compact 17/1, https://www.compact.nl/articles/its-nothing-personal-or-is-it/, 2017.

About the authors

S.E.J. Jeurissen MSc is a senior consultant at KPMG Data & Analytics. She specializes in Enterprise Data Management. She advises organizations on the design and implementation of policies, rules and procedures needed for the effective management of (un)structured data.

Ir. P.C. Scherpenhuijsen is CTO of Indica, he cofounded the company back in 2013. Indica develops and delivers software solutions to automatically correlate, process, index and archive structured and unstructured data from digital and non-digital sources. The software delivers all information by using amongst others patented correlation algorithms, entity recognition and natural language processing (NLP).