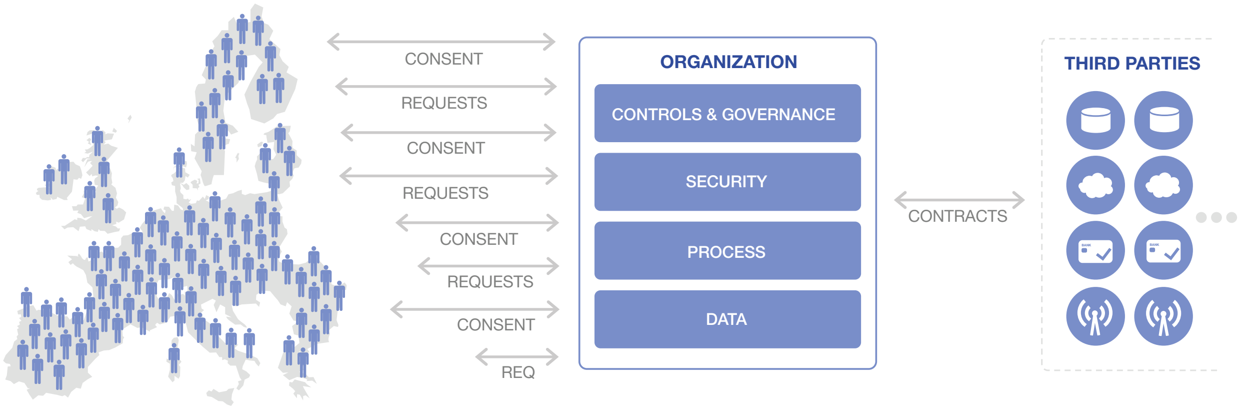
MarkLogic®

EU GENERAL DATA PROTECTION REGULATION - THE PATH TO COMPLIANCE

MARKLOGIC WHITE PAPER • NOVEMBER 2016

This white paper addresses one of the key regulations that top executives have on their agenda: EU GDPR. What is it? What are the key data challenges and the fastest way to compliance? These and many other questions are answered in our publication.



EXECUTIVE SUMMARY

The regulatory landscape is becoming more and more complex with new mandates being introduced across various jurisdictions. One of the most significant regulations affecting all companies in possession of European citizens' personal data is the EU General Data Protection Regulation (EU GDPR) coming into effect in May 2018.

In this white paper we address some of the key issues related to this new regulation and provide practical guidance into solving them in a timely and cost-effective way, including:

- What the EU GDPR is and how it impacts your company
- Data readiness checklist
- Challenges and risks of achieving compliance
- A step-by-step guide to compliance: MarkLogic can help
- How to leverage new technology in your data management process for new opportunities

WHAT THE EU GDPR IS, AND HOW IT IMPACTS YOUR COMPANY

The EU General Data Protection Regulation (GDPR) defines the rights of EU citizens (also known as Data Subjects) around the privacy and protection of their personal data. The regulation also specifies the requirements of any organisation or individual that is responsible for determining the purposes of storage and processing of this data (also known as Data Controllers). It will be the Data Controllers who will be held responsible for any breaches of this regulation and upon whom fines will be potentially levied.

Any organisation that is responsible for employee, customer, patient, citizen, etc. data will be subject to this compliance. The EU GDPR comes into effect in May 2018. *Failure to comply with this regulation can result in fines of up to 4% of global revenue.*

The extent of the regulation is far reaching, and achieving compliance will be complex, especially for large organisations. Reducing the risk of this regulation will require proactive planning as well as the ability to ensure compliance over time.

For example, it is critical to institute a framework whereby an organisation can quickly respond to individual (data subject) requests. These requests may include:

- A written notice requiring you to not use their personal data to make automated decisions, asking about reasoning made about an automated decision, or asking you to reconsider an outcome of an automated decision
- A request on the data processing patterns and whether their data will be given to other organisations or persons
- Requesting a copy of their personal data
- Withdrawing consent for usage of their personal data in specific cases (with certain restrictions) or for direct marketing
- Requesting the correction of inaccurate personal data

Each of these requests has time-frames in which they

DATA READINESS CHECKLIST

The 9 questions you need to answer to ensure data readiness for EU GDPR:

- á Do you know what data in your organisation contains personal data, including unstructured sources such as documents?
- á Do you know how you use or process personal data and for what purpose?
- á Do you know what consent was given (or later retracted) for which purposes, for this personal data?
- á Do you know what security controls are in place for personal data, and whether this data is deleted or anonymised appropriately once the consented usage has finalised?
- á Can you respond to regulatory requests about the usage and storage of personal data, in the timeframes required by the GDPR?
- á Can you scale your ability to respond to potentially thousands of concurrent requests (e.g., as part of a class action lawsuit)?
- á Do you store personal data in multiple languages?
- á Are you treating personal data as a valuable asset to enhance customer experience and potentially develop new products?

must be addressed. For example, the organisation has 28 days to respond to the request to withdraw consent for direct marketing.

And there is a further complication.

Whilst the regulation itself is finalised, each member state must interpret the regulation as appropriate for their local laws and decide upon recommendations and enforcement practices. In the UK, for example, the authority responsible for this is the Information Commissioner's Office (ICO). The ICO have stated that it will take many more months to complete this recommendation. Until that time, it will be hard for any organisation to know exactly what steps they must take in order to comply with the regulation.

CHALLENGES IN ACHIEVING COMPLIANCE

The regulation has two major mandates any organisation has to adhere to:

- Effectively identify and store with all security controls the personal data in their internal systems
- Provide immediate responses to EU citizens about the usage of their personal data or it's deletion from internal systems

In order to comply with the above imperatives, organisations must know what personal data they store or process and for what purpose, what consent has been granted, and what security and controls are in place. Moreover, organisations must then know how this data links together to pertain to individual data subjects, in order to respond to an individual request.

The larger the organisation, and the more disparate the systems that deal with this data, the more challenging it will be to gather this information and respond to requests.

USAGE AND STORAGE OF PERSONAL DATA

The regulation mandates that personal data can only be stored or processed for the use and purpose for which the data subject has given consent. For example, archiving of this data for non-regulatory or legal reasons may not be permitted. Another example is around general analytics: data that contains personal data may need to be anonymised before use. If a data breach occurs, and it was discovered that personal data was stored or processed in a non-compliant fashion, the responsible organisations will potentially face legal action.

It is therefore necessary for organisations to identify what data they store and process, and make decisions on its continued existence and usage – and adapt as needed based on future changes to regulations. However, this step alone presents organisations with an enormous data processing activity.

TIMELY RESPONSES TO EU CITIZENS ABOUT THEIR PERSONAL DATA

In order to respond to requests about the usage of personal data, or to respond to the withdrawal of consent for a particular usage of that data, organisations must first know what personal data is pertinent to the request. In order to understand this, they must know what Data Subject (citizen, customer, employee, etc.) pertains to what data. For example, a customer may cancel a contract with an organisation, whilst still having other contracts active. Organisations must know specifically what data elements refer to that customer and that specific contract and its associated processes. Once all relevant data is identified, the organisation can then begin the surgical removal of that data set.

Organisations that choose not to organise personal data at this level of detail present themselves with the following risks:

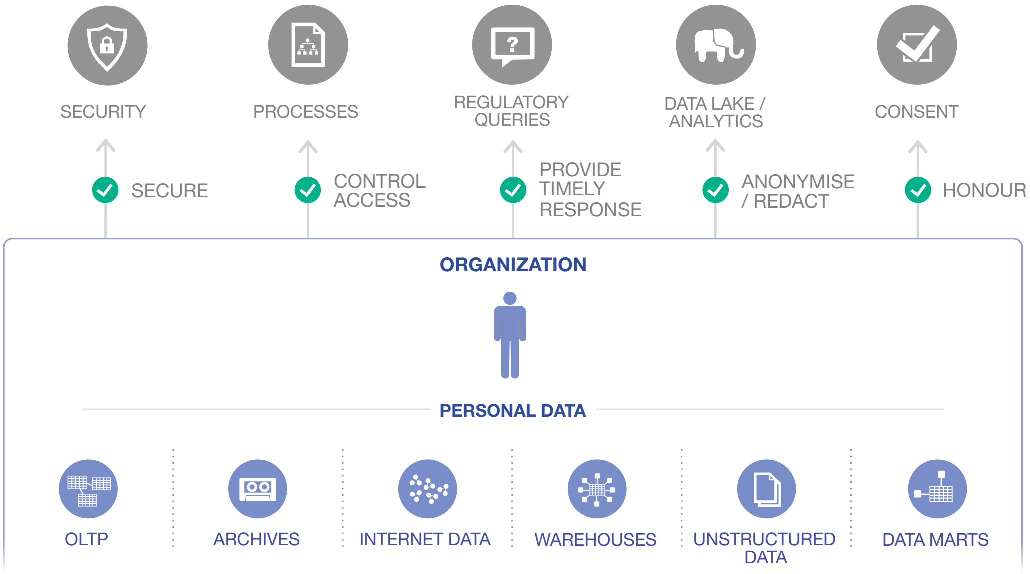
- The effort involved in responding to a request may make it hard – or impossible – to complete the action within the required timeframe. This is because even if an organisation knows what systems contain personal data (and for what business usages those systems are used), the extra effort required to drill down into each of these systems and identify a specific individual can be an expensive exercise.

- The ad hoc nature of responding as-and-when, may mean each request is timely and expensive, making long-term costs greater than the effort in organising the information in the first place.

- The co-ordination of large groups of individuals making requests at the same time may mean that even if a single request could be processed adequately, dealing with dozens – or even hundreds – of concurrent requests becomes impossible. It is also likely such co-ordinated groups will have stronger legal representation capable of class-action lawsuits.

Organising personal data at the level of the data subject can be performed manually or with automated support. Manual approaches will typically require intensive human effort, that will (for even modest data estates) be prohibitively expensive. Worse, whatever manual output is generated, will still only produce a static result that would require ongoing effort to maintain.

Organisations could instead look towards traditional large vendors for tooling to aid in a (semi) automated effort. These tools, however, will typically have two shortcomings, when it comes to application of the GDPR compliance:



> “ Whilst the regulation itself is finalised, each member state must interpret the regulation as appropriate for their local laws and decide upon recommendations and enforcement practices.

- They will suit either structured data (e.g. relational databases), or unstructured data (e.g. documents), but not both. Organisations will have personal data across their entire data estate, and will require a solution that can manage information that spans multiple data types
- They will lack the ability to annotate these data sources with the rich metadata required to understand and analyse risk and exposure, and to respond to regulatory requests in a timely manner

A STEP-BY-STEP GUIDE TO ACHIEVING COMPLIANCE

STEP 1. IDENTIFICATION OF PERSONAL DATA

The first step is understanding what personal data your organisation has. This data could be stored in database systems, applications (both on-premise, and cloud based), or documents (Word, PDF, etc.). Examples of applications could include HR systems, websites (which may store cookie data, or other online identifiers), CRM systems, etc.

Organisations should first identify all possible systems (whether owned and managed internally, cloud-based, or owned and managed by 3rd parties, but for which the organisation is responsible for the data stored in the system) in which personal data may exist.

Once the relevant systems have been identified, organisations can then inspect the data stored in these systems and apply matching rules (which may range from pattern matching to identify telephone numbers, through to natural language processing to understand the complex usage of language in various documents) to identify any personal data. Once personal data has been identified, its nature, broader context and provenance should be saved to a system that allows this information to be easily retrieved and queried.

STEP 2. DOCUMENTING USAGE AND PURPOSE OF PERSONAL DATA

Once a data source has been identified as containing personal data, the next step is to document the purpose and usage of that data.

For certain data sets, knowledge about – for example – what business process consumes that data set, or what retention policy is applied to it, already exists in some other systems (e.g. in a business process management system, or in an asset management system). This data can be imported into a database technology to provide a single consolidated view of the personal data, annotated with all associated details (referred to as “metadata”). Running queries against this combined knowledge base reduces complexity and cost compared to alternative approaches such as trying to federate the queries across multiple different systems.

For other data sets, and for certain details, this knowledge may not be known or currently captured. If so, organisations will require a manual process of documenting this knowledge. Automating this process will require a new technological solution, which has data discovery capabilities that enable the categorisation of data in order to uncover some of these details.

Importantly, the GDPR is still evolving in terms of understanding what knowledge about the personal data is required. Organisations require the ability to have a flexible way of recording this personal metadata to satisfy the possibility of a new attribute being required later.

> “The GDPR is still evolving in terms of understanding what knowledge about the personal data is required. Organisations require the ability to have a flexible way of recording this personal metadata to satisfy the possibility of a new attribute being required later.

STEP 3. ZOOMING IN ON INDIVIDUALS

The previous steps illustrate how personal data can be identified and annotated with information about its usage and purpose. However, there is a third step that is also very useful in being able to respond to a regulatory request about an individual: the identification and joining of data about that individual across multiple systems. For example, a customer may have data stored in an e-commerce application, a CRM system, and a marketing system; and, if an organisation has gone through an acquisition or merger, that data may be duplicated across both originating organisations' systems.

Matching rules can be applied that link multiple data sets together into entities. Data matching is rarely a precise exercise, as – for example – addresses may have changed, making the confidence two entities are the same less than 100%. So, it may be necessary to store the confidence of an association. Moreover, the discovery of this inconsistency in address (for example) can be of value to the organization, i.e. in helping uncover inaccurate data.

Once entities have been identified across data sets, queries that provide responses to regulatory requests can be refined to the level of an individual, providing the most accurate results possible. Without this extra insight, a query can only be made to the level of granularity of the type of data in each data set.

A final point on this step, is that this linking of data sets into entities can be performed iteratively. In the same way that the annotation of data must be flexible to future requirements, so the linking of data sets should be performed and refined as necessary to meet regulatory requirements for compliance in the most timely and cost-effective way. For example, it might be quick and easy to identify 80% of customers by name, and address, or some identifier. For the remaining 20% the exercise may be delayed until there is a need to reconcile the data sets, e.g. in response to a regulatory request regarding that individual.

STEP 4. ANONYMISING AND REMOVING PERSONAL DATA

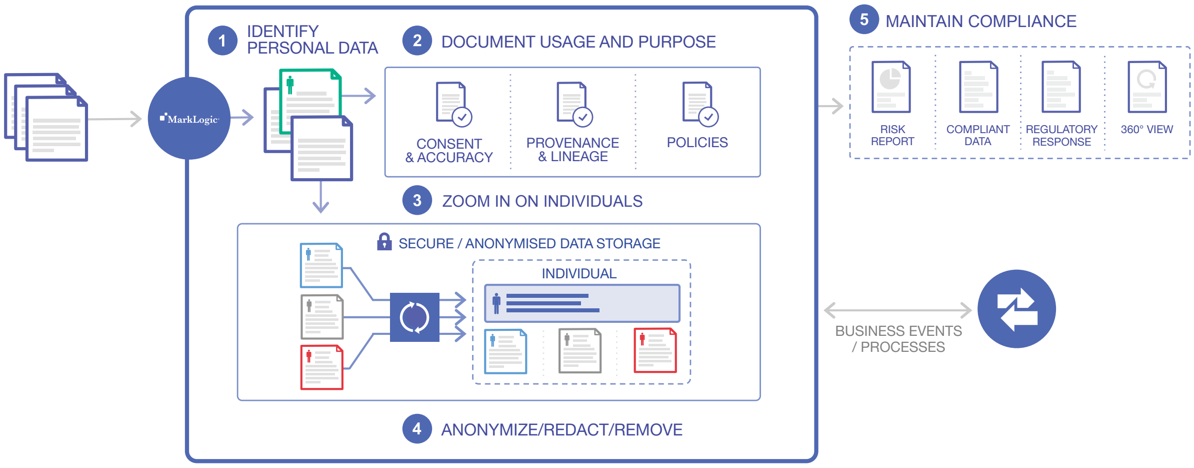
Now that the organisation understands what personal data relates to what Data Subjects, and how that personal data is being used – and what consent has been granted, the organisation can now query this knowledge base to identify if personal data is being used or stored outside of this granted consent.

At this point, organisations can choose to anonymise (or redact) the data and/or to remove the data in the case they have no legitimate use for the data. Finally, in the case the data has partial consent (e.g. for one usage, but not another), organisations can apply dynamic redaction/anonymisation or filtering, depending on who is requesting the data.

STEP 5. MAINTAINING COMPLIANCE

The final step is to ask questions about the personal data in order to satisfy a request. Rich search and query capabilities are required over the collated knowledge set of the organisation's personal data in order to collect and present the necessary data to respond to the request. Questions can be asked of the data itself, of the details it has been annotated with, or both.

For example, an individual may complain that they are receiving direct marketing; it might be that the user initially consented to such marketing, but only for a particular length of time (maybe over a particular season, or during a specific offer). The organization will want to be able to query details about that individual (perhaps only using an e-mail address), targeted to a specific marketing data-set, and pulling back what consent was given, for what campaign, and for what time period. Importantly, once the appropriate consent



has been identified, the organisation would need to update the system to potentially flag that the user has tacitly withdrawn consent. At this point, appropriate processes would then need to be initiated to update the systems so as to prevent further direct marketing.

Again, due to the evolving nature of the GDPR, the solution must be flexible in the queries and questions it can support.

Importantly, the ability to gain this deep insight into your customers, employees, patients, etc., can be used for more than just responding to regulatory requests. These same capabilities can be used to drive deeper insight into these individuals, and thus turn this system into a platform of business value.

THE SOLUTION REQUIREMENTS

The main requirements for a technological solution to support the above steps, and expedite compliance with this new regulation, are:

The ability to extract data from systems and analyse it to identify personal data (whether this data is highly structured, or unstructured; and potentially supporting multiple languages)

- The ability to record the nature of the personal data and annotate it with a rich set of details, the specifics of which may change over time
- The ability to match and link the personal data into Data Subject entities
- The ability to react to data changes across the estate so that this knowledge base is kept current
- The ability to both:
    - Co-ordinate between source systems in order to remove data or update data as consent is granted or retracted
    - Store personal data where the source system isn't capable of providing necessary redaction/ anonymisation or filters (in order to only present personal data to the appropriate processes and users), or doesn't provide the necessary security to store the personal data
- The ability to provide rich query and search across the personal data and its metadata in order to respond to requests and business events (such as consent being withdrawn), and be flexible enough to support new queries and questions as the regulation evolves, and other business use cases for the solution emerge
- Provide enterprise capabilities such as hardened security, high availability, and even the ability to run on premise or in the cloud (as IT requirements dictate), since this solution will be touching the organisation's most important data

MARKLOGIC'S CAPABILITIES

MarkLogic® is an operational and transactional Enterprise NoSQL database platform that is designed to integrate, store, manage, and search more data than ever before. As a multi-model database, it can handle any kind of data, structured or unstructured – and unlike traditional databases, it does not require the extensive upfront schema development or ETL processes that cost organisations significant time and money.

The MarkLogic database platform provides a Google*-like multi-lingual search across enterprise data, and enables not only discovery questions to be asked about the data, but also rich complex queries suitable for gathering necessary information to respond to a regulatory request.

MarkLogic can identify personal data in your data, and can annotate and enrich this data, allowing the recording of how and why the data is consumed, who is responsible for it, and any other attribute that would be required to respond to requests.

MarkLogic is a security certified platform, ensuring access to this information is restricted to appropriate individuals, and can anonymise and redact the data as required.

MarkLogic integrates with business processes so that it can be part of a broader GDPR solution, and can initiate processes to delete or modify data in it source systems.

MarkLogic can be installed on-premise, or in the cloud; and, it is operationally cost effective to run as it typically requires a fraction of administrative overhead compared to other database platforms.

HOW TO LEVERAGE COMPLIANCE REQUIREMENTS FOR NEW OPPORTUNITIES

The EU GDPR is an extensive and potentially disruptive regulation that can have huge costs to organisations who fail to comply. However, there are benefits that can be realised through adopting the solution outlined in this whitepaper.

Managing data in this new way – supported by new, innovative technology – provides an opportunity to achieve a complete view of the key individuals that comprise or interact with that organisation.

This information can gain insight into the touch-points of that individual, what business processes interact with that individual, and what is known (and subsequently, what should be known – but isn't) about that individual. This insight can give organisations the ability to improve customer and employee satisfaction and trust; it can also reduce costs by leading to more efficient business processes, improve decision-making by using more complete and accurate data, and align business units through a unified view of customer/employee/citizen/etc.

Having a holistic view of data is a powerful mechanism to analyse customer behavioural patterns—helping to enhance the stickiness of the current product offering and shaping a path to innovative product development.

The MarkLogic approach enables organisations to turn their compliance effort into a value proposition: providing an opportunity to improve customer/employee/citizen trust and satisfaction, helping to reduce operational costs, and delivering deeper insight into customer profiles and the products they are most interested in.

MORE INFORMATION

Plan for Success with High-Stakes Data Projects
marklogic.com/resources/plan-success-high-stakes-data-projects/

Beyond Relational
marklogic.com/resources/beyond-relational/

Contact us to further explore our solutions at www.marklogic.com

MARKLOGIC CORPORATION
999 Skyway Road, Suite 200 San Carlos, CA 94070
+1 650 655 2300 | +1 877 992 8885 | www.marklogic.com | sales@marklogic.com

MarkLogic®

999 Skyway Road, Suite 200 San Carlos, CA 94070
+1 650 655 2300 | +1 877 992 8885
www.marklogic.com | sales@marklogic.com