Spotlight

Spotlight Paper by Bloor
Author **Philip Howard**
Publish date **August 2016**

The data management implications of GDPR

> …the adoption of data profiling, data quality, data lineage, data masking, test data management, data governance, and a 360° view of the customer are all likely to be affected by GDPR compliance. In addition, there are particular types of data processing that will be impacted, such as archival and analytics.

Author **Philip Howard**

Executive summary

The General Data Protection Regulation was passed in April 2016 and is due to come into effect in May 2018. It is intended to provide a consistent approach to data privacy and protection across the whole of the EU, in other words, any company that processes or controls EU citizen data, regardless of where they are based. Put briefly, GDPR extends existing data protection legislation by giving individuals additional rights over how their data may be used by companies that collect and process that data. It also imposes various obligations on those companies.

This paper will describe the requirements of GDPR – both for EU and non-EU organisations – in some detail. However, the main emphasis in this paper is on the implications of GDPR compliance from a data management perspective. There are a number of these. In particular, the adoption of data profiling, data quality, data lineage, data masking, test data management, data governance, and a 360° view of the customer are all likely to be affected by GDPR compliance. In addition, there are particular types of data processing that will be impacted, such as archival and analytics. The use of age verification software will increase and the deployment of cookies on websites will need to change. In most cases it will be particular aspects of the legislation that will drive data management requirements and this paper is therefore organised to reflect these cause and effect relationships.

> “…the main emphasis in this paper is on the implications of GDPR compliance from a data management perspective.”

The legislation

GDPR does not, as some have claimed, mean that consumers “own” the data that companies gather about them. What it does mean is that they have rights over how that data may be used and processed.

Unlike the Data Protection Directive which preceded it, the General Data Protection Regulation, is an EU-wide law that is intended to provide conformity across the EU. This is the difference between a *“regulation”* and a *“directive”,* the latter being open to wide-scale interpretation but the former being limited to different national implementation in only limited respects. For example, the main focus of GDPR is that users have rights over how their may be processed. It might be supposed that the regulation requires "explicit" consent but technically GDPR states that the purpose for which consent is being given must be clear (unambiguous), while consent must be constituted by an affirmative action. Consent in this case is similar to the sort of consent that must be provided before undergoing a medical procedure – informed, explicit and current – and, in most jurisdictions it is recognised that minors cannot give such consent. However, the age at which maturity is deemed to be reached varies from country to country and the GDPR allows for individual countries to have different rules for the age at which consent is required. Age verification software is likely to become increasingly important to support GDPR compliance. Suppliers are required to make “reasonable efforts” with respect to establishing the age of customers.

GDPR does not, as some have claimed, mean that consumers “own” the data that companies gather about them. What it does mean is that they have rights over how that data may be used and processed. As mentioned in the previous paragraph, apart from specific exceptions which will be discussed later, they must be asked for their consent before their data may be processed. The legislation is specific in this context. Firstly, it is an opt-in environment rather than opt-out. Secondly, this may not be hidden in the small print: what customers are opting for must be clear and unambiguous. Thirdly, the legislation states that it “may not be valid” for companies offering a product or service to make this conditional on opting in – best practice would be to assume that it is not valid – at least until after some case law has been established. One immediate consequence of these conditions is with respect to websites that use cookies. Companies collecting user behaviour through cookies will either have to give users the choice of not opting for cookies or they will have to capture non-consent data and filter out the cookies of those who have not opted in.

In this context it is worth providing some figures. A recent survey conducted by dataIQ in conjunction with Experian found that 16% of consumers were willing to share their data if they had trust in the company asking for it, a further 33% would share their data if it was explained why it was needed, and a further 49% would prefer not to share their data unless it was really necessary. This survey was specifically with respect to marketing and we would expect to see lower figures for more abstruse IT functions (archiving, DevOps and so forth) that do not directly impact on consumers.

When it comes to consent, GDPR refers to “processing” of your data and there are various ways in which processing may be achieved. Technically, the act of storing data on disk is a form of processing, but so are archival, testing and development, business intelligence and analytics and similar functions. We will consider these in more detail in due course.

In addition to the retained (and slightly amended) rights from the Data Protection Act (such as the right to certain minimum information, the right of access to personal information, the right to object and the right to rectification, erasure or blocking) data subjects now have the right to be forgotten (an extension to the right to erasure) and the right to data portability. While each of these rights may be actioned by the customer at any time it is the giving and withdrawal of consent that will have the most far reaching consequences.

Apart from consent and the rights over data processing, there are a number of other areas with which GDPR is concerned. These are:

1. The extension of the definition of personal information. For example, email and IP addresses, and genetic data, were not considered private under the Data Protection Directive

but they are under GDPR. It is possible that consumers will allow the use of some personal information but not others. For example, a customer might consent to the use of data combined with their email address but not if it is combined with their physical address or phone (mobile or otherwise) number.

2. Companies over a certain size are required to appoint a Data Protection Officer (DPO). While employed by the enterprise the DPO is effectively mandated to work on behalf of the customer rather than the organisation per se. The word used in the regulation is that he or she must be "independent".
3. Companies collecting data (with consent) have a duty of care towards the data. If that company monetizes the data by selling it on to some third party, then the originating organisation has a duty of care to ensure that the data is not used for purposes beyond the boundaries of the consent provided. What this means is that either or both companies can be fined for a breach.
4. Breaches are to be reported *"without undue delay and, where feasible, within 72 hours after being discovered"*. Delays of more than 72 hours need to be justified and breaches need to be reported not just to the regulator but also the person or persons whose data has been impacted by the breach. The exception with respect to data subjects is if the company can demonstrate that appropriate security measures – such as encryption (or masking, referred to as pseudonymisation in the regulation) – have been implemented, so that the data would be rendered unintelligible to anyone accessing it.
5. On the subject of fines, these can be up to 4% of global revenues or 20 million euros, whichever is greater. Organisations that have appropriate compliance procedures in place will not necessarily escape fines but the fact that they can prove that they have taken appropriate steps to comply with their obligations "will be taken into account".

Non-EU countries

GDPR covers both businesses in the EU and non-EU businesses that conduct business in the EU, either by offering goods or services to EU data subjects, or by monitoring the behaviour of EU data subjects. If a company falls into any of those categories, it must comply with the requirements of GDPR.

The UK

The UK will be a non-EU country at some point in the future but it will still be in the EU when GDPR comes into force. We believe the UK will adopt the GDPR or something very, very close to it due to its desire to trade closely with the single market. Firstly, UK companies will all be scrambling to comply before 2018, and if the UK government then decides to hit them with a UK version that is significantly different it will probably cost a lot to comply with both laws, even just from a technical standpoint, and will cause a lot of consternation. Secondly, the UK government has so much legislation to unravel that it will be looking for easy options wherever possible, and thirdly, having significantly different data protection laws in the UK will make it even more difficult to trade within Europe. For political reasons the UK government may choose to implement the requirements of GDPR by extending the existing Data Protection Act but we do not believe that its provisions will differ significantly from those of GDPR.

> **"We believe the UK will adopt the GDPR or something very, very close to it due to its desire to trade closely with the single market."**

Processing options and data management

There are a variety of different ways in which user data may be processed, which need to be examined, along with their data management implications, in turn. However, before doing so, it is worth considering how and where you store relevant consent information. To begin with, it should be noted that there are exceptions to the requirement for consent. These exceptions relate to public interest, research, historical and statistical purposes and are typically subject to anonymisation.

For commercial organisations, the possible situations are:

- There is no consent to store data. This will be relatively rare as most consumer engagements require storage (see next) at least for a temporary period.
- There is consent to store data and consent for it to be used for some specified purpose or purposes.

> “While non-storage is trivial from a data management perspective consent data is another issue. In effect, you need to check whether you have consent before you process any individual’s data for any relevant purpose. How is this to be achieved?”

While non-storage is trivial from a data management perspective consent data is another issue. In effect, you need to check whether you have consent before you process any individual’s data for any relevant purpose. How is this to be achieved?

The simple option is to extend all the applications that process personal data. Companies would need to add an extra column or columns (or the equivalent in non-relational environments) to each relevant customer database within their organisation, and modify their applications to process this consent data. This may be simple but it will not be practical in most instances. An alternative would be to rely on third party vendors. Most companies processing personal data will be doing this using master data management, customer engagement solutions, business intelligence, data integration, data quality, data preparation, analytics and archiving tools that are provided by suppliers from within the IT community. It would be reasonable to expect vendors of products such as these to build in the ability to recognise consent data and process (or not) records accordingly. The original company would still need to modify their database tables but at least would not have to worry about their applications.

Another option is to build a consent management application through which all requests for customer data are filtered. This would work rather like dynamic data masking in that calls to a particular database are intercepted and processed to check that the application has authorisation to process this particular data. In practice, and subject to some issues around data masking that are discussed below, this is a sensible approach that could be combined directly with dynamic data masking, although tools that offer this capability would probably have to be modified by their vendors. This is because such products tend to be rules-based and because they would need to be extended to support consent data. A more generic capability without any dynamic data masking is also a possibility.

The big problem with any of the vendor-oriented solutions outlined is that none of these suppliers currently offer what has been suggested. They can be expected to in the future because it will clearly be competitively advantageous to do so, but these solutions do not exist today. The only exception is within DevOps and test data management (discussed later) where the use of synthetic data bypasses the need for actual data.

Storage

The simplest processing issue is whether you have permission to store a consumer’s data. In practice, you might think this was implicit. For example, tweeting or posting data to a Facebook page logically implies consent to that data being stored as otherwise it could not be posted. However, GDPR requires that consent be unambiguous and social media sites will have to make this clear. For children under the age of consent, GDPR states that *“processing shall be lawful only if and to the extent that consent* *is given or authorised by the holder of parental responsibility over the child"*. The significant words are "given or authorised" so a child without authorisation could potentially buy a game on the Internet without involving their parent (for example, using Bitcoins) as long as they download the game immediately and that after the download the supplier does not retain or further process any of the personal data related to that purchase (including the IP address). Gaining consent or authorisation is probably an easier option.

Similarly, ordering a product that needs to be physically delivered, means that address details need to be recorded. Moreover, subject to returns policies, product guarantees and so forth, one assumes that that information would need to be retained for whatever period is defined. However, there is an expiration date in such cases. Companies will need to decide – the regulation does not make this clear – at what point companies need to ask for consent to use your data beyond this expiration date. This could be when products are bought or it could be after the returns or guarantee period expires. In any case, the consumer has the right to withdraw consent at any time and companies must enable that capability.

A refusal to give consent for data storage at the outset is not technically challenging. In the case of web browsing you (the company) either don't employ cookies or you throw the cookie away. Neither of these is difficult even though it may require a change to your web processing.

A removal of consent is another matter entirely. In order to comply with this, companies will need to know where the relevant personal data has been stored. Depending on the extent of the consents originally provided this may be in many places: transactional systems, data marts and warehouses, data lakes, archives and back-ups. It will be necessary to identify where the original data entered the organisation and how the data flowed through the organisation from there. In practical terms this will typically mean using one or more of the following:

- Data profiling to identify related data across multiple data sources.
- Data lineage (which often relies on data profiling) to establish the data flow for this particular data and where it has landed throughout the organisation.
- Single view of the customer. Once a client has decided to withdraw consent in one area it is very likely that they will do this more broadly. In order to support such withdrawals of consent – and also to allow customers to correct or update their details – it will be useful (arguably, necessary) to have a 360º view of that customer that includes all relevant details about him or her. Indeed, it will be useful if this view has been extended to include other details about the customer (call centre notes, emails, social media comment and so on) as these may all need to be removed from your systems. Note that this argument also applies to consumers that want to review their data so that they can correct any mistakes.

There is a potential side benefit here. It is estimated that data quality for personal data deteriorates at between 1% and 1.5% per month. This is because people move house, die, get married and change their names, buy new mobile phones, get issued with new credit cards, and so forth. If customers could be "trained" to update their data whenever these things occurred that would improve data quality in general. According to the dataIQ/ Experian survey mentioned above, more than half of respondents recognise this is as their responsibility and this figure could no doubt be increased if sufficient attention was paid to this issue.

Finally, with respect to storage, back-ups are going to be a major issue. Suppose a user withdraws consent. Then all relevant data must be removed and

> **"It is estimated that data quality for personal data deteriorates at between 1% and 1.5% per month."**

> **…certain types of database are made impractical by GDPR. The fact that consent may be revoked has direct implications here.**

this will often include back-up data as well as live data. There are two issues: firstly, identifying the back-ups that hold data that needs to be removed and, secondly, the actual process of deletion. The first of these issues is not insuperable. Probably simplest would be to have consent data time-stamped so that you know when consent was valid and you can identify back-ups taken during this period. This is likely to be best practice in any case, certainly from an auditing perspective. The second issue, however, is more difficult. How do you remove data from a back-up, while retaining the use of the back-up? Simply put: you can't. Or, at least, not without wholesale changes to database and application technology, which simply isn't going to happen. The only alternative is that you have to create new back-ups every time someone changes their mind about consent. This is not practical. It would require huge investments in infrastructure in order to meet everyday performance requirements. We imagine that an approach that could be called *"eventual compliance"* will be considered satisfactory, provided that *"eventual"* means as soon as is practical.

Data masking and anonymisation (the regulation refers to pseudonymisation) is a potentially contentious area. While masked data clearly isn't relevant to one-to-one marketing it is potentially important in other scenarios such as allowing personal data to be used for analytics based on aggregated data or for test data management or for (medical) research purposes. The question is whether it is permissible to process and use someone's data if it is masked (we are here talking about static data masking rather than the dynamic data masking referred to above) or anonymised, even if you don't have consent?

The answer is not clear cut. The problem is twofold. Firstly, you have to store data before you can mask it. If you don't have consent for the former you can't use the data. Secondly, while anonymised data is not considered to be personal data, data masking is never perfect. For example, your largest customer will still be your largest customer even if his name is masked. Also, it is sometimes necessary with data masking that you can get back to the original unmasked data and there are algorithms and techniques that explicitly support this capability. In practice, the applicability of masking is likely to remain in doubt until and if it is settled in court. We would recommend that companies using data masking have explicit processes in place to prove that masking algorithms either cannot or have not been reversed: detailed auditing will be required.

While on the subject of practicality, certain types of database are made impractical by GDPR. The fact that consent may be revoked has direct implications here. Specifically, append-only databases – and many NoSQL databases are append-only – may be in trouble. In these databases, when data is deleted the data is flagged as *"deleted"* but is not actually removed from disk and, at least in theory, the data can still actually be accessed. In order to truly remove data, you have to use a workaround, typically by *"dropping"* an object. Unfortunately, objects tend to be large constructs (tables, partitions and so forth), which means that you would have to drop the whole customer table and append the new customer table without the removed consents. This is clearly not feasible. Therefore, if you want to use Hadoop as a data lake, for example, and use it to store personal data, you will need to select a distribution that employs conventional read/write capabilities as opposed to the native append-only structure of HDFS (Hadoop Distributed File System).

Archival

Archival is an extension to data storage for historic data. Much the same considerations apply to archives as apply to conventional storage except that we would expect consent to be required for data to be archived for use in the future. From a vendor perspective, information lifecycle management products could easily be extended to include consent-based deletion and correction. It should be noted that the regulation explicitly allows the archival of data if it is in the public interest and provided that processes are in place to prevent the identification of data subjects (**see box**).

Analytics and business intelligence

This is the area where customer data is most likely to be required – for one-to-one and other personalised marketing – and where consumers are most likely to see potential benefits for themselves. On the other hand, it is the most likely to be annoying: pop-up ads for something you already bought somewhere else, repeated emails about the same things, incorrect spelling of one's name, and so forth. It is also worth noting that trust is a fundamental issue here: if you lose public trust people will opt out, and once lost trust is very difficult to regain.

There are several points worth making. Firstly, with respect to under-age participation on social media. The important point is that relevant tweets or pages may be stored but they may not be used in any other way, because their owners cannot give consent for that. This implies that companies doing social media analytics will have to filter out all under-age data before doing such analyses. Twitter itself, along with other social media sites, will also have to obtain consent from adult users before their tweets can be used for anything other than the use for which they were originally intended. Further, Twitter will have a duty of care to see that third parties do not perform social media analytics against non-consenting data.

A second group of people that will be impacted by GDPR are (citizen) data scientists. These people will typically build algorithms or analytic applications based on an analysis of customer behaviour. However, these individuals are often not authorised to see relevant personal information even if consent has been given. In the latter case, and because of GDPR, tools for data scientists, such as data preparation platforms, will need to have an understanding of consent as well as data masking (**see box**) to obfuscate personal data that they are not authorised to see.

A further point with respect to analytics is that there is nothing to stop companies stripping out personal data – assuming there is no consent – and retaining what is left. For example, it might be useful to know how customers are traversing your website even if you can't keep the personal data attached to that browsing behaviour. You can usefully aggregate the data without it being in any way personal.

Consent also raises issues about the treatment of data lakes. If a data lake is treated as a place where you put data that you might want to analyse at some point in the future, then consent will need to be part of that picture and, depending on your implementation, you might need to go back to the consumer for consent, when and if you decide to use their data (often this will be social media) for analysis. Note also the previous discussion on append-only databases.

Finally, a lot of marketing analytics is about segmenting customers. Customers providing or not providing consent represent segments in their own right, and the removal of a non-consenting proportion of customers will skew analytic results.

DevOps

DevOps needs to use production data for testing or, at least, test data derived from production data. Briefly, there are three ways to achieve this: you can take a copy – which may be a virtual copy – of your production data; you can take a sample (subset) of your production data or you can generate synthetic data. There are advantages and disadvantages associated with each of these approaches and this is not the place to discuss these, with the exception of the fact that copied or sampled data needs to be masked while synthetic data, because it is not real, does not. Thus, from the perspective of GDPR both copying and sampling production data will be impacted because consent will be required, while synthetic test data generation will not. In so far as the former is concerned, we would expect DevOps to gain consumer consent much less frequently than for other types of processing, to the extent that it may actually not be practical to use real data, because it is not truly representative of the production data as a whole. Indeed, one could argue that the self-selecting nature of consent means that the data will, by definition, be non-representative and would skew results, as discussed in the previous section.

> “…because of GDPR, tools for data scientists, such as data preparation platforms, will need to have an understanding of consent as well as data masking.”

Data management

A number of data management disciplines have been mentioned during the course of this paper and it is pertinent to discuss each of these individually. Relevant technologies include:

- Data profiling. This will be required to identify personal data that is subject to consent. It may also be used as a preliminary discipline to support data cleansing and governance, and to enable synthetic test data generation. More advanced data profiling tools have the ability to identify relationships that exist across data sources and this will be important in establishing data lineage and in enabling data preparation platforms where data is to be merged across data sources. Similar considerations apply to the implementation of master data management and (extended) 360° views of the customer.
- Master data management and 360° views will be important in supporting the rights of consumers in correcting or deleting relevant data.
- Data lineage, while not a product category in its own right, will be vital in getting an understanding of how and where personal data is used and re-used across the organisation. Data lineage is typically enabled by both data profiling and data integration tools. Some products extend pure data lineage by allowing you to see where the data has been masked so that the security and privacy of the data can be monitored as the data flows through organisational processes.
- Static data masking is typically applied to test data but it is also appropriate for use with data preparation platforms where data scientists and business analysts are bringing data together from disparate sources.
- Dynamic data masking, which is often aligned with database activity monitoring, has historically been used to prevent unauthorised access to data, typically for legacy applications where role-based access control has not been implemented within the application. We believe that such products could easily be extended to provide consent-based access control.
- Data integration has a role to play since the transformation engines in such products can easily be used to support the conditional (based on consent) movement of data. Similar considerations apply where APIs are used.
- Information lifecycle management products, which are typically used in archiving scenarios, need to have consent-based options built into them.
- For test data, choices have historically been between sampled or copied versions of production data which are then masked, or synthetic test data management. The adoption of GDPR will favour the latter.

> “Master data management and 360° views will be important in supporting the rights of consumers in correcting or deleting relevant data.”

Conclusion

Companies needing to comply with GDPR will have to decide whether to adopt a Napoleonic or Anglo-Saxon approach to this new law. The former, which is prevalent in the EU, is that nothing is allowed unless it is specifically authorised. The latter, common to the United States, United Kingdom and elsewhere, takes the stance that everything is permitted unless it is specifically banned. This is likely to be reflected not just in how IT vendors support GDPR but also in how different legislatures adopt the principles of GDPR. In practice, the EU is likely to take a Napoleonic approach so it will probably be sensible for organisations to take a similar line.

More generally, GDPR throws up more questions than it does answers. For example, there are some areas where a literal reading of the regulation suggests completely impractical solutions (for instance, deleting data from back-ups). More generally, there are few if any IT vendors that have ready-made solutions to implementing GDPR. Certainly there are suppliers of data management tools that can help you to enable consent-based processing but ERP, CRM, social media and other providers are singularly lacking when it comes to consent-based processing. This is unusual, it is more often the case that technology already exists to enable new regulations, but GDPR goes deep into the heart of the applications that are needed to run business and which will now need to be re-examined and re-structured.

FURTHER INFORMATION

Further information about this subject is available from *www.bloorresearch.com/update/2300*

About the author

PHILIP HOWARD

Research Director / Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), and dining out.

Bloor overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations, and in 2014 celebrated its 25th anniversary. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter 'noise' and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent 25 years distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

Copyright and disclaimer

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**