Bloor

Spotlight

Spotlight Paper by Bloor
Authors **Daniel Howard and Philip Howard**
Publish date **January 2020**

Sensitive Data Management

“In this paper, we shine a light on the technologies and capabilities needed to discover, secure and manage the sensitive data within your environment.”

Introduction

We should start by defining "sensitive data". While most people will think of this as being about personal data and its protection, this is by no means the only type of corporate information that is sensitive. For example, company financial information is sensitive, especially if you are a listed company. Similarly, intellectual property details need to be protected and so do things such as marketing plans, new product launch details and so on. None of this particularly impacts on how you manage sensitive data, and in the remainder of this paper we will typically be discussing personal data, but it needs to be borne in mind that sensitive data also has other connotations.

Most organisations are storing sensitive data, but often do not know where it is or which data it is. Moreover, some companies are storing data that they do not realise is sensitive, which may be either directly or indirectly sensitive. In the latter case we are referring to data that can be used to re-identify anonymised data. For example, researchers at the University of Texas took de-identified data released by Netflix and by comparing this data with movie reviews on a third party web site, and knowing about as little as only two movies a user had reviewed, including the precise rating and the date of rating (give or take three days), this allowed for a 68% re-identification success.

Most organisations also have a (possibly loose) understanding that managing sensitive data is important. We have already discussed what sensitive data is: by managing that data we refer to identifying, protecting, and retiring said data while at the same time providing curated access to that data. Essentially, it must be safe, but it must also be useful.

Moreover, all of this must comply with the various data protection regulations and mandates you are subject to or will be subject to in future, most prominently (but certainly not exclusively) GDPR. The most obvious consequences for non-compliance are by this point well known, consisting in GDPR's case of an extremely hefty fine, but in addition to that there are also significant reputational costs associated with non-compliance, and in particular with data breaches where the data in question is both sensitive and unprotected. Quite frankly, the public mood has turned on this sort of thing, with organisations that allow personal data to be leaked suffering severely in the public eye. Also note that data breaches are rapidly becoming a matter of "when, not if". Making the assumption that you will never suffer a data breach is simply untenable at this point in time, no matter how good your security practices are. Thus, managing your sensitive data in the fashion we describe is not just recommended but essential, both for avoiding fines and for maintaining your organisation's reputation and thus your userbase.

As we have just mentioned, most executives are (now) onboard with the idea of managing sensitive data. However, many are left in the dark as to how exactly to go about it. This is supported by a CIO Watercooler survey, which found that managing sensitive data was the number one data-related issue for CIOs. Hence, this paper, in which we intend to shine a light on the technologies and capabilities needed to discover, secure and manage the sensitive data within your environment.

"Most organisations are storing sensitive data, but often do not know where it is or which data it is."

Sensitive data discovery

> Your primary requirements will be to a) find the sensitive data in your system and b) expose it so that it can be classified and anonymised.

Like many other terms used within IT, data discovery is used in more than one context. In particular, some business intelligence vendors have historically described their capabilities as providing data discovery, though we would categorise this as – more accurately, we feel – insight discovery. However, it is also used in a more literal sense, to discover what data you have where, how data in different data sources is related and what those relationships are, and whether there are any dependencies that exist between data elements, regardless of whether those elements exist within a single database or across multiple, potentially heterogeneous, data sources.

Your primary requirements will be to a) find the sensitive data in your system and b) expose it so that it can be classified and anonymised (see next section). Given our introduction above, it should be clear that the first of these steps, that of identifying the sensitive data present within and throughout your organisation, is not trivial. Moreover, it should be obvious that you must first identify your sensitive data before you can classify, anonymise or indeed do anything else with it. This makes sensitive data discovery an extremely important initial step.

There are a number of different methods for identifying sensitive data, and the type(s) of tool required will depend on the type of data in question. For databases, a common type of product used will be a data profiling tool. However, these are general-purpose offerings with discovering sensitive data just one supported task among many. Alternatively, there are specialist products that have a greater focus on sensitive data discovery. These typically derive either from the database security space, the data management market, or have historically focused on unstructured data. A typical process, would be to sift through your data, assign a probability to each piece as to whether it is sensitive based on predefined criteria, then present that data to you as potentially sensitive if that probability exceeds a defined threshold, ready to be tagged or otherwise classified at your discretion. However, many of these tools – there are exceptions – do not work for either spreadsheets or text files or, for that matter, with (many of the) NoSQL databases. In these cases, you may need specialised capabilities such as spreadsheet governance tools.

In database environments, the primary differentiators between data discovery tools (at least in terms of data discovery as a specific capability) come in terms of how such data is identified. For instance, it is very common to offer simplistic "column name" matching that will try to determine if data is sensitive based on the name of the column in which it is stored, for example if the field contains the word "address". Anyone who has ever worked with a large database will immediately realise the problem: it relies on your database having well thought out, descriptive and accurate column names, and for any new columns to be similarly well named. This is hardly common, which makes column matching something of a non-starter when used as a standalone capability. Thankfully, more sophisticated techniques are available, serving as the aforementioned differentiators. There are products, for example, which will match by examining column data and metadata, or by performing code introspection. Note also that you will often want to deploy several of these methods at once (even column name matching) in order to ensure a high degree of coverage. In addition, machine learning is very well suited to this process, owing to the fact that it is both probabilistic and highly repetitive. Its presence can also serve as a differentiator, as can leverage of primary/foreign key relationships or analysis of user queries to the database. Distance measures such as when a postal code appears close to the name of a town are also appropriate, and these can can also be used with unstructured data.

Finally, false positives need to be minimised. For example, when scanning email servers an email that simply says "Philip, thanks for xyz" does not represent sensitive data as, on its own, it cannot be used to identify anyone.

Data Masking, Encryption and Tokenisation

Once you've found your sensitive data, you will need a way to protect it, by which we mean anonymising it. Your options, in essence, are to mask it (replace it with generated, non-sensitive data), obfuscate it (either wholly or partly) or encrypt (either reversibly or irreversibly) or tokenise it. Which of these options you prefer, and/or require, will likely depend on a) your security requirements, b) your use case(s) and c) your performance needs.

Data masking is in most cases a highly secure option, because it completely replaces your existing data. However, this comes at the obvious cost of losing access to the original data. The point is that your masked data can be generated using masking algorithms to retain relevant properties of the original data, while at the same time being entirely fake. For example, you may need your masked data to retain the same structure as the data it is replacing (for example, a credit card number), for it to fall in the same range or have roughly the same value, or for certain parts of it to be calculated deterministically (such as a credit check number). Essentially, data masking is useful when you don't care about the data itself, but only the underlying properties of the data.

There are two types of masking: static masking and dynamic masking. The former masks the data at rest and the latter in-flight. Or, perhaps more accurately, the distinction is that static data masking replaces real data with masked data, whereas dynamic masking does not replace anything but directs unauthorised access to a masked view of the data. Static data masking is most commonly used, or should be used, in non-production environments such as development and testing, as well as in analytics and data science (where it is less common than it should be). Here the ability to preserve the structure of the data you are masking is important, maintaining such things as referential integrity or ensuring that you use consistent masking algorithms across multiple datasets that you may want to join. An alternative approach, at least within DevOps, is to leverage synthetic data generation, which sidesteps the issue.

Dynamic data masking does what its name suggests: masks the data at run-time, typically by intercepting SQL requests and then masking the results. This allows for the possibility that some users may be allowed to see more unmasked data than others. It is typically used in production environments.

Two further options are encryption and tokenisation. The differences, including advantages and disadvantages, related to each of these techniques, as well as data masking, are summarised in ***Table 1***.

Table 1:

Encryption	Tokenisation	Masking
Uses an algorithm and key to transform data into an encrypted form	Random generation of a token value with the mapping between them stored in a database	Uses a choice of algorithms to replace real values with look-alikes. Can be applied either to static data or dynamically
Usually requires encryption key to decrypt but you can use an encrypted mapping table to improve performance	Requires access to token database	Cannot usually be unmasked though reversible masking algorithms are available
All three techniques nowadays support both structured and unstructured data		
Also used for entire files	Commonly used with structured data fields such as credit card or social security numbers	Most appropriate where it is the structure of the data that is important rather than actual values
Good for exchanging data with third parties that have encryption key	Difficult to exchange data because of access to token database	Easy to exchange masked data but third parties cannot see original data
Format preserving encryption is possible but less secure	Format can always be preserved without threatening security	Format is always preserved but possibility of re-identification
Original data may leave the organisation but only in encrypted format	Original data never leaves the organisation	Original data never leaves the organisation
All three techniques can co-exist. For example, you might have active customer data masked, inactive customer data tokenised, and highly critical data might be encrypted. This allows you to optimise for scale (see below), as well as providing flexibility.		
Scales well with just a small encryption key	Does not scale well because token database increases in size. Conversely, requires less compute power to process	Scales best because nothing additional to store

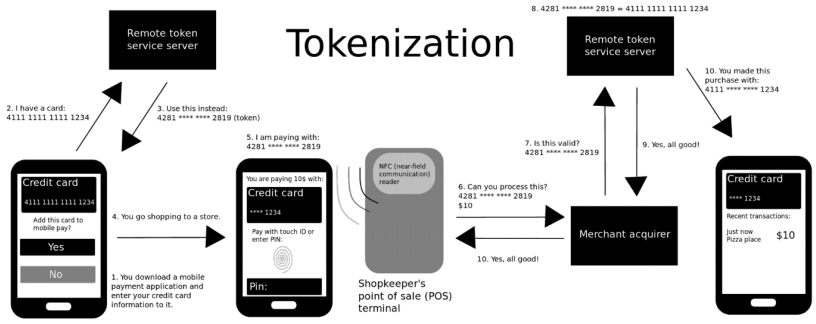
Encryption, and in particular irreversible encryption, will likely form the backbone of your anonymisation, since unlike masking it allows (highly regulated) access to the underlying data, a necessary evil for most use cases. Format preserving encryption (FPE), which is reversible, also exists and does what it says: the encrypted data has the same format as the original data. However, reversible encryption is significantly less secure (and as a result, in a significant number of countries/situations it does not meet compliance regulations) because it can, in principle and sometimes in practice, be reversed by a malicious party. FPE can be useful, usually when dealing with automated systems (for example, test automation) that need a specific data format but don't care about the contents, or where you have something like a credit card number, for which some users need to see the last four digits but others should not see anything at all. Format-preserving encryption is relatively new and is currently fairly niche, though it poses a threat to tokenisation (see below). Homomorphic encryption is another option, which allows computation on encrypted data, generating an encrypted result which, when decrypted, matches the result of the operations as if they had been performed on the unencrypted data. Historically this method has had performance issues, but there are companies working on development in this area.

The way that tokenisation works is illustrated (albeit in a simplified version) in ***Figure 1*** (courtesy of Wikipedia).

Unlike data masking, both encryption and tokenisation are cryptographic security methods. The major difference between them, at least historically, has been that tokenisation is format preserving while traditional encryption is not. This is a function of the fact that tokenisation is non-mathematical. However, the introduction of FPE changes the picture.

Finally, obfuscation is a technique that is essentially a very crude version of masking or encryption. Think censorship with a blue pen: part or all of a particular data element is simply blanked out.

Figure 1:



Risk

For some types of sensitive data, particularly non-personal data that is not subject to the sort of compliance restrictions that personal data is, it may be appropriate to use a risk score to determine priorities when it comes to de-sensitising data. This is commonplace, for example, in spreadsheet governance products. In some instances, for instance medical research, this also applies to personal data, where you need to assess the risk of re-identification.

Data Governance, Policy Management and Monitoring

As with many digital transformations, there are two core issues that you need to deal with when managing your sensitive data. The first is managing the sensitive data you have now; and the second is managing the sensitive data that is continually flowing into your organisation. In order to achieve the latter, you must take the processes of discovery and anonymisation detailed above and put them into practice on a continual basis. This requires data governance and, more specifically, policy management. You will also want to be able to control who has access to your sensitive data (whether anonymised or not) and will therefore need authentication and authorisation capabilities in place. This is also provided by data governance, in conjunction with tools such as Active Directory and LDAP. Finally, you will want to be able to monitor the sensitive data within your organisation, so that you can see how your sensitive data is being accessed, by whom, and when. In particular, you will want a security monitoring capability that can detect and report anomalous – and what could be malicious – behaviour. This can provide you with advance warning of a data leak, possibly allowing you to prevent it or, at least, respond to it quickly and appropriately.

The main point to bear in mind here is that this must all be done continuously. When new data enters your system, you should be automatically determining if it is sensitive, anonymising it if it is, and applying access rules as appropriate. This is most easily done via (automated) policy management, which in turn will normally be provided by a data governance solution. This should allow you to create policies that mandate the automatic discovery and anonymisation of any sensitive data as it enters your system, thus allowing you to manage incoming sensitive data indefinitely and thereby make sure that your organisation doesn't relapse into noncompliance.

It's also important to note that this kind of governance is crucial even over multiple systems and environments. What's more, you will want to be able to manage the sensitive data in these environments consistently, regardless of location. Given that different environments will often use different tooling, different data formats, and so on, this can be difficult. It is also precisely the problem that the open source ODPi Egeria project (part of the Linux Foundation) was created to resolve. This is accomplished by applying common metadata standards, with the first release providing a "single view" of metadata across your organisation, along with federating queries across and synchronising metadata between metadata repositories. Although it's clear that Egeria is still in its early days (the initial version is available on GitHub), this is already a good start towards the goal of enabling one-time governance over all of your systems.

Consent

Wikipedia has an out-of-date definition for consent management that states that *"consent management is a system, process or set of policies for allowing consumers and patients to determine what health information they are willing to permit their various care providers to access. It enables patients and consumers to affirm their participation in e-health initiatives and to establish consent directives to determine who will have access to their protected health information, for what purpose and under what circumstances. Consent management supports the dynamic creation, management and enforcement of consumer, organizational and jurisdictional privacy policies."* While this is accurate in so far as HIPAA compliance is concerned, it should be extended to legislation such as GDPR and CCPA. However, the basis principles are correct. Unfortunately, there is also a definition going around that reads something like this: *"consent management*

"There are two core issues that you need to deal with when managing your sensitive data. The first is managing the sensitive data you have now; and the second is managing the sensitive data that is continually flowing into your organisation."

is a process which allows websites to meet the EU regulatory requirements regarding consent collection", and there is a class of products calling themselves consent management platforms which aim to fulfil this function. While a valid activity for companies wishing to ensure that email campaigns, website access and so on are compliant, this is only a subset of the requirements for consent.

More generally, consent doesn't directly affect either the discovery or the protection of sensitive data but, for personal data, it does define what needs to be protected and in what contexts. User consent therefore needs to be collected and managed and it is therefore an element of governance, where relevant policies and processes – as per the Wikipedia definition – should be in place to determine how personal data is classified and, therefore, how and when it should be protected.

Data Archival and Retirement

A key aspect of GDPR is that you should only hold onto sensitive data as long as you still have a specific, well defined use for it. Once that use has been exhausted, you should remove that data from your live system, either archiving it or deleting it entirely. Similarly, GDPR enshrines the rights of users to have their data retired (in other words, removed) from your system on request, the so-called "right to be forgotten". In any case, regular archival of (anonymised) sensitive data is good practice for minimising the damage caused by data leaks (and remember that data leaks are a when, not an if).

A standard pattern for data retirement has emerged. All ingested data is equipped with an expiration date as it enters your system. Once this date passes, the data will be flagged for archival, and on some regular time period (once a month, say) an automated process runs that archives any flagged data. Similarly, archived data may be marked for deletion on a future date after it is archived, and a similar process is run to delete any archived data so marked. However, many applications do not allow the deletion of recorded data. For example, financial or human resources data needs to be retained because of data integrity constraints, where removal would create mismatches against reported/real data. In these instances, it is appropriate to tokenise the sensitive part of the relevant records but without deletion of those records. Whichever approach is appropriate, the relevant rules can be applied and enforced using Information Lifecycle Management (ILM) policies that you've defined. In addition, when a request is received to retire data, you need to be able to flag this appropriately and on an ad hoc basis. Note that you will not necessarily want (or be required) to act upon this request immediately, as doing so would a) invalid your database backups and b) does not account for the fact that your users are liable to change their minds. Archiving first and deleting or tokenising later neatly solves both of these issues.

Note also that the process of archival implies removal of the data from source systems and that this therefore only applies to databases where this is possible: many NoSQL databases are append-only databases where it is not possible/practical to delete data. Such databases should never be used to store sensitive data unless they have specific capabilities enabled to support sensitive data retirement.

"Regular archival of (anonymised) sensitive data is good practice for minimising the damage caused by data leaks (and remember that data leaks are a when, not an if)."

Conclusion

It is not easy to manage sensitive data across an entire organisation. Not only must you discover, protect, monitor and retire the sensitive data within your systems, you must do so continually and in perpetuity. That said, sensitive data management is a vital component of your IT infrastructure that you cannot afford to ignore. The alternative is it fall foul of GDPR and other regulations and attempt to survive while dealing with hefty fines and substantial losses of reputation. Remember that data breaches are not something that *"happens to other people"*: if you are an organisation, your (sensitive) data is always at risk. Therefore, it must be protected.

In this paper, we have attempted to elucidate the most important capabilities a comprehensive solution for sensitive data management should have. Having said that, it is not an exhaustive list. However, it is a good starting point, and any solution that can cover the areas described above will be likewise.

"Sensitive data management is a vital component of your IT infrastructure that you cannot afford to ignore."

FURTHER INFORMATION
Further information about this subject is available from
www.bloorresearch.com/update/2536

About the authors

DANIEL HOWARD
Analyst

Daniel started in the IT industry relatively recently, in only 2014. Following the completion of his Masters in Mathematics at the University of Bath, he started working as a developer and tester at IPL (now part of Civica Group). His work there included all manner of software and web development and testing, usually in an Agile environment and usually to a high standard, including a stint working at an 'innovation lab' at Nationwide.

In the summer of 2016, Daniel's father, Philip Howard, approached him with a piece of work that he thought would be enriched by the development and testing experience that Daniel could bring to the table. Shortly afterward, Daniel left IPL to work for Bloor Research as a researcher and the rest (so far, at least) is history.

Daniel primarily (although by no means exclusively) works alongside his father, providing technical expertise, insight and the 'on-the-ground' perspective of a (former) developer, in the form of both verbal explanation and written articles. His area of research is principally DevOps, where his previous experience can be put to the most use, but he is increasingly branching into related areas.

Outside of work, Daniel enjoys latin and ballroom dancing, skiing, cooking and playing the guitar.

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

Established 30 years ago, Bloor has become one of Europe's leading independent IT research, analysis and consultancy firms.

Bloor is widely respected for providing actionable strategic insight through its innovative independent technology research, advisory and consulting services. Bloor assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

Underpinning Bloor's whole ideology is that digital and business transformation isn't a serial 'one and done'. Being a Mutable Business™ in a state of permanent reinvention; evolving business models, people and resources with technology is the key to securing long term survival.

Bloor Consulting

Bloor Consulting is here to guide your organisation on its journey to be a Mutable Business™ Our senior-level consultants leverage in-house research, best practice data and vast industry experience to meet and exceed our client's expectations.

Bloor offers a range of packaged or custom consulting services. Short, or long-term, consulting engagements. Depending on where you are in your journey, these are just some of the areas we can advise your organisation.

For a full list see ***bloorresearch.com/consulting-service/***

Disruption & Change
Bloor can help you assess and evaluate opportunities and threats to your business (and marketplace) from the emergence of new organisations and technologies. If you are looking to the future and considering how best to innovate, Bloor can help fully evaluate the technologies and opportunities arising, and create an adoption roadmap, risk assessment and change plan to support your innovation.

Strategy
Bloor's unique blend of technology competency, delivery, business, and commercial experience enables us to support your strategic planning, ensuring that realistic expectations are set, and risks are appropriately managed. As well as ensuring the impact of change is fully reflected in your business plans.

Business Creation & review
Bloor can assist you to create a practical vision of how developing technologies and emerging working practices will create opportunities for your business. We can introduce you to organisations developing the innovation as well as transferring our knowledge and experience to your business.

Cybersecurity
According to the latest World Economic Forum poll, cyber-attacks are seen in 2019 as the most pressing risk for CEOs in Europe and North America, including six of the ten largest economies in the world. Cybersecurity risk is a whole board agenda item and our experts can advise you on the latest and best ways to protect your organisation

Leadership
We can undertake a review of your Mutable Business™ journey so far and help your team to understand the impact and full consequences of the change needed. Our team can work with you to create a vision of the future resulting from the change.

Advisory and Review
If your business has a high dependence on external 'partners' to deliver major service contracts, are you concerned in ensuring you can control their cost and change processes? We can provide advisory and review services and support, to ensure you can keep their costs, and your agenda, on track.

Copyright and disclaimer

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

tel: **+44 (0)1494 291 992**
web: **www.Bloorresearch.com**
email: **info@Bloorresearch.com**