



COA¹ Paper

Information Classification

Introduction

The aim of this paper is to demonstrate how information classification can be used to enhance security within a Collaboration Oriented Architecture. While there has been a great deal of work done on information classification, this has tended to be bespoke to any and every organization that has used it. Furthermore, classification has traditionally only been used on a very small percentage of information assets. In the new de-perimeterised world this is no longer a realistic option.

Problem

Within organizations, information is growing at an alarming rate and looks to continue to do so for the foreseeable future. In a de-perimeterised environment this problem is worse. Data is going to be created and utilized by more than just employees, with partners, suppliers, consultants and customers all having a hand in your data.

Ultimately, the classification of information ultimately corresponds to the level of protection afforded it and so consistency is required, not only within an organization but also across organizations living in the de-perimeterised business environment.

Current information classification systems are designed for specialists to use and subsequently only a very small percentage of information is labelled. The problem is worsened when you look at some of the other factors which need to be taken into account when looking at protecting information.

Why Classify Information?

There are a number of reasons why organizations pursue an information classification program, including:

1. Control over access to sensitive or confidential information.
2. Protection of sensitive or confidential information.
3. Simplifying the discovery of sensitive or confidential information.

Before embarking on an information classification program, it is worth analysing why you are doing it. Is it for regulatory reasons? Is it as part of a corporate governance initiative? Will the information be used to help in information storage and protection or not? Quite often it is a mixture of all three, but this too is important to your understanding of how to evaluate the possible solutions.

The Value of Data Varies Over Time

Classifying a document is not a static problem; it varies over time. For example, an initial document may not contain any sensitive information until its third draft, when acquisition targets are named. While this process can occur relatively quickly over a matter of days, the process can take a great deal longer, for example state secrets may become de-classified after 50 years. Any classification mechanism needs to take time into account.

¹ The Collaboration Oriented Architecture paper and associated COA Framework paper are available online from the Jericho Forum publications Web page at <http://www.opengroup.org/jericho/publications.htm>

Other Considerations on Value

Data has different values to different people. It's misuse or loss can result in differing consequences to the organisation "owning" the data and those impacted. In everyday life we look to open and efficient markets to set prices. However there is no open and efficient market for most of the company confidential information that businesses deal in - it is rarely freely tradable² - so its "value" is hard to assign.

The key phrase here is "to different people". We deal in preventing the harmful side effects of misuse of information. The most important element of any risk specification is the person(s) or thing(s) which would be impacted. This may be an individual, a group of individuals, an organization, a market, an industry, a country, etc. So in assessing the value of data, we need to understand "who we are protecting". Impact level may be viewed as secondary to this and should be evaluated in this context. What is serious to one person may be trivial to another - hence, one person may categorise information as "Highly Sensitive" while another may regard it as only "Normal Business".

A further consideration on "value" is organizational obligations - for example, to a regulator (protecting a market), or to a Data Protection registrar (protecting individuals), or to a client organization (protecting proprietary data). In a collaborative environment, a key question is who owns these obligations, and what obligations can be delegated to whom?

Aggregation of data also significantly impacts its value (see later in this paper).

The Risk to Data Varies by Location / Geography

Information being accessed via a cyber-cafe probably needs to be treated and protected differently to that which is being accessed on a desktop inside head-office. It may be that some information which has been classified as 'Highly Sensitive' is not allowed to be accessed from a cyber-cafe, or even from a smart phone. For example good practice in local health authorities is not to allow health records on smart phones in case they are lost (even though they tend to have encryption and other adequate security measures implemented on them).

While the classification of the data remains constant no matter where it is, the risks are different and a policy engine working with classification to protect information needs to take geography and / or location into account.

Classification versus Action

Classification is a starting point from which actions can then occur. There is no implied definitive action or actions based on the classification. However, various policies and technologies can be applied to various classifications. For example:

- Data outside the organizations may be encrypted. This can be achieved in a number of different ways:
 - Laptops could have full disk encryption.
 - Individual files could be encrypted, with a key shared by all employees.
 - Individual files may be subject to electronic digital rights management (eDRM) to enable sharing of the information with 3rd parties in a secure manner.
- Less important data could be protected in a different manner to that which is more important.
 - Files could be stored on more secure storage, for example, mirrored or RAID5.
 - Backups could happen twice a day.
 - Archive copies could be taken.
 - The information could be synchronously replicated to a second and / or tertiary site.

² There is an underground economy that deals in individuals information such as credit card numbers, bank details and even usernames and passwords. For this type of information there is a market value – however it is small compared to the reputational and individual damage that can be done if the information is lost. For example, credit card numbers can be bought for only a few cents and bank account details for a few hundred dollars.

Consistent Information Classification is Hard

If something has been classified, especially when it is done by an individual, then consistency becomes an issue. What one person thinks of as being ‘Highly Sensitive’ another might think is just ‘Sensitive’ and a third might think is ‘Public’. For example, a CEO probably deals in ‘Highly Sensitive’ information all the time whereas a sales manager doesn’t. For the sales manager, they may classify an ongoing deal as ‘Highly Sensitive’ when the correct labelling would be ‘Sensitive’.

The problem becomes harder still when working with a Collaboration Oriented Architecture. When there are multiple parties specifying classification there needs to be agreement on how the classification is carried out and how the information is subsequently handled. A government’s ‘Sensitive’ information probably requires more rigour applied to it than a widget manufacturer!

Key issues for a Collaboration Oriented Architecture

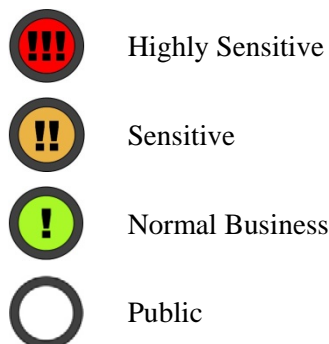
Thus the key question for anyone designing an information classification system in a COA framework is; “what levels of classification do you require and how do you protect the information at each level”. Any checking must extend beyond Operation Systems checks on file-level security to (potentially) include being able to assess the security status of all applications that are used in the transaction that modifies this file, as well as (potentially) checking that unwanted application are not running at the time the file is modified.³ Such checking may be one-time or may need to be continuous, depending on the type of collaboration and the process involved in the transaction.

Classification Levels

While there are a number of different information classification systems around, a simpler ‘less is more’ approach is recommended so as to reduce confusion. To this end, the proposal put forward by the G8, commonly called the ‘Traffic Light Proposal’, makes a great deal of sense as it is simple to explain and therefore relatively simple to follow. There are four levels:

- Red: Highly Sensitive. Personal for named recipients only.
- Amber: Sensitive. Named Groups.
- Green: Normal Business. Business community wide. (This is the default classification for organizations.)
- White: Public. Public distribution, unlimited control.

Tagging classification using colours has to be augmented with another visual clue such as an icon. The following icons are proposed:



It should be noted that there are a number of different classification schemes available. Some have more levels and others less levels. It is a worthwhile task to map any scheme you are currently using to the G8 traffic light protocol, as this will make sharing the information with other companies possible without either side having to change their classification scheme. For example, in BS17799 there are five levels:

BS17799	G8 Traffic Light
Public Documents	White (Public)
Internal Use Only	Green (Normal Business)

³ See Clark-Wilson integrity model for specifying and analyzing an integrity policy for a computing system - <http://www2.computer.org/portal/web/csdl/doi/10.1109/SP.1987.10001>

Proprietary	Green (Normal Business)
Highly Confidential	Amber (Sensitive)
Top Secret	Red (Highly Sensitive)

Of course, whether 'Proprietary' maps to 'Sensitive' or 'Highly Sensitive' is entirely up to the user organisation and depends on what type of information placed in that category.

OASIS lists a large number of classification standards. More information can be found at: <http://xml.coverpages.org/classification.html>

Personally Identifiable Information

While "classification level" is important, it is worth drawing out one other category of information - Personally Identifiable Information (PII). Protection of PII has become a global issue, after wide experience of the consequences of a number of data leaks by companies and governments alike. Legislation to protect PII is in place in the United States of America and is in progress in Europe.

The EU directive 95/46/EC defines PII as:

Article 2a: 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

Examples of this include, but are not limited to:

- Full name
- National identification number
- Telephone number
- Street address
- E-mail address
- IP address (in some cases)
- Vehicle registration plate number
- Driver's license number
- Face, fingerprints, or handwriting
- Credit card numbers
- Digital identity

This information is often classed as 'Sensitive', although the consequences of it being lost or leaked means that it should probably be treated as 'Highly Sensitive'.

Aggregation

When dealing in electronic information and records it is important to look at the effect of aggregation. Using PII as an example, if you were to lose a single record then the impact is not nearly so great as if you were to lose a million. Electronic systems make it very easy to aggregate records and to copy them around as easily as if it was just one.

Classification of information may apply to entire data sets. For example, for a customer database with PII information, it is not just the database that is 'Highly Sensitive' but each record too, should it be copied into other documents.

When dealing with partners and suppliers in a Collaborative Oriented Architecture it is important to agree the ground rules on what is and isn't acceptable. For example, it might be acceptable that the partner can share information with their partners, but only one record at a time rather than giving access to the whole database.

When it comes to data loss and data breach notification, understanding exactly how many records are involved becomes essential. Losing one person's personal information may be less serious than losing many, though the impact to that individual could be quite devastating. The organization, however, will be much more concerned over information losses of sufficient size to attract regulatory

sanction or public awareness/disapproval. If the organization can prove that the loss is for a limited subset of records then notification can be limited. If not, then every customer may have to be notified.

Information Ownership

In a Collaborative Oriented Architecture it is easy to lose sight of who owns the information and who assigned any classification to it. In general the creator of the information owns the information. They may classify it, in which case this classification should be respected by everyone who subsequently handles that information. If a sensitivity is assigned, then the recipient must not re-classify the information at a lower sensitivity.

Information Classification Taxonomy

While the classification levels outlined above are useful, organizations are also interested in classifying information to provide additional information - for example, is this information related to 'Mergers and Acquisitions' (M&A) or is it just a social event? For this reason, organizations may want to look at a classification taxonomy, especially when sharing information in a Collaboration Oriented Architectures. In that way, design documents can be separated from legal ones, and so forth.

There are a number of different methods for creating taxonomies as well as a number of standard ones - for example the Dewey Decimal Classification system for library books. It is worth noting that creating taxonomies can be time consuming and difficult and in many cases the law of diminishing returns is rapidly reached, but the people defining the taxonomy may never reach the end. When creating an information classification taxonomy, it should be limited in both time and scope so that it is created efficiently. It should also be noted, that this is not an "IT-only" endeavour, but rather one which involves the whole organization, including lines of business, legal, finance and of course, IT.

Defining and agreeing a number of high level categories is worthwhile, and these can then be used by individuals as well as automated classification tools.

Automated Classification

With the ever increasing quantity of electronic information and records, the need for automatic classification is also growing. Automation can be used to address the changing requirements based on time and geography as well as content.

The rules for classification of information need to be defined. There are some patterns in information, such as credit card numbers which are universal and easy to define. For others, keyword, phrases or regular expressions are used and it is relatively simple to use the same definition across multiple classification engines. However, there are also statistical analysis tools which result in a classification being given. These tend to be proprietary and the definitions almost impossible to share among the classification engines in order to ensure consistency.

If automatic classification is used then it is worth considering archiving the definitions and ensuring that the version used is also stored with the classification. In that way it will be possible to tell how a particular document was classified and therefore the actions that happened because of it.

Multiple Classifications

When classifying information into a taxonomy there are probabilities attached, usually based on statistical analysis. In this case, there can then be multiple classifications with attached probabilities. While it is usually only the most probable classification that is acted upon, it may be useful to take into account the second or third probability, especially if they are relatively close. For example, if a piece of information is classified as 'Social.. 0.78', 'M&A... 0.77' then the second classification would result in completely different handling of the information than the first.

Challenges to the industry

Current classification methods and mechanisms are proprietary. While it is relatively simple to define criteria for some information, such as credit card information, most company confidential information is unique to the company that creates it. A common method of defining classification rules is required to enable the sharing of classification information. Similarly there needs to be a standard method to tag information with the classification, including its sensitivity, taxonomy and probability.

For Open Collaborative Architectures to share information effectively, classification information needs to be shared. In order for that to occur, the key vendors in the space need to agree an open specification / standard they will all use by which such classifications and definitions can be shared.

The Way Forward

The ability to consistently classify information at all points in its life cycle and across the entire IT infrastructure is critical. If the information cannot be classified correctly then it will not be able to be managed appropriately. Static classification of information by the information owner is not workable in today's global environment and so consistent automation is also required.