To Backup or not to Backup

Abstract

The title actually alludes to the current availability of almost unlimited IT storage space and the consequent rush to use it. So we "archive" in it everything and more, without selection nor thoughts. But this can be quite costly, as some recent IT incidents have proven. It is then necessary to evaluate which kind of information to process in IT form, what to discard immediately and what to archive for longer time, independently of storage availability.

In very few years IT has revolutionized the way in which we do business and create and communicate information. This revolution happened and is still happening too fast. We, as human, have not yet learned to use and manage the immensely powerful tools given to us by IT and we are loosing control of our own information.

The title of this article is a small provocation about the meaning of producing and keeping, or storing for the future, all the information that we manage in the IT context. To understand if what we are doing now is right, wrong or partially right and wrong, we need to start from the beginning, that is understanding how information is produced, managed and stored in the IT context.

From the latest Incident ...

Before delving into our considerations on information management in the IT world, we can understand that something is not right from the, by now very numerous, IT incidents. One of the latest and best known incident is the so called "Sony Hack" [1] (November 2014) in which a huge amount of company, private and sensitive information of the Sony Corporation has been made public. Besides stealing information for obvious financial gain, like stealing full credit card numbers, or for economical damage, like releasing on the web a new film, the fact that other company, private and personal information is made public can have a wide range of negative

consequences. In the case of the Sony Hack, private and personal information of the employees has been made public which has caused both the embarrassment of the employees and the starting of legal action against the company by the employees themselves for the personal damage occurred. Also sensitive emails and comments of some managers have been made public, with very bad consequences for the reputation of the company.

To the current trends

Data leak or data exposure is the current trend, this is also officially recognised by recent laws and regulations that impose to companies the public disclosure of incidents when data[1] have been made public. The purpose of this is to inform and alert the data owner of the risks due to the public leakage of information. If your credit card information has been lost by or stolen from a company, it is very important that you are informed so that you can block the card before fraudulent charges are made. At the same time, the disclosure of the incident makes public some information about the damaged people and companies, but it is considered to be a minor risk and damage (one could call it a collateral damage) with respect to the benefits of informing the damaged parties. Moreover, by giving a bad reputation to the company affected by the data leakage, it has the direct effect of forcing the company to improve its security measures to try prevent it to happen again.

This is to be also matched with the current social media madness, at least from this point of view, where voluntary public disclosure of personal information is seen as "the right thing to do." This of course just adds to the trend, since here we are interested in what companies do and not so much what the general public does. But it is certain that the current social media frenziness has wide impacts on how companies behave, since we are the same persons when we are at work and at home.

Of course all of this would not be possible if the IT technology would not offer us, by now at discount prices, the means to produce, manage and disseminate all this information. As of today social media are pervasive and easy to use, and storage is cheap so we can put in it all what we want, from the pictures of our vacations, to an out of place comment, to all arguments, right and wrong, during our job.

---

1 There is actually a difference between the words "data" and "information", here we will not be too careful about this difference and we use "data" usually when we consider the technical IT aspects and "information" usually when we consider the human aspects.

So it is worth to look at all this from the point of view of a company, starting from what the technology allows us to do, to how we should use it and the risks in between.

Social media, data, storage and information

Let' s start by considering the social media point of view. The IT technology of social media offers us the possibility to communicate, exchange information, fast, often in real time, in all possible kind of situations and media. One can post a comment, view what's going on, be always up-to-date with the current developments in whatever one is, should or could be interested. This is a huge amount of data, part of which is information and most of which is just replica of information, which is produced, distributed and stored.

Indeed, apart from the ubiquity of social media, one important aspect is that in the previous list the word “deleted” is missing. With very few exceptions of some services built on purpose to “rapidly forget”, information is produced and never deleted, it is there to stay for ever. This is seen as a major distinctive characteristic of these services since up to not too long ago the capability of storing large amount of information was expensive if not almost impossible. Indeed, just to mention one of the most famous, Facebook was created with the purpose of storing, in one place, large amount of information about school mates usually difficult to archive, and is so doing, to give the possibility to access in a easy and fast way all this information.

So on one side it is easy to produce information, with all the “social” technology at our disposal, from smart-phones to digital cameras, and recently also watches and glasses. And more often information is just grabbed as images, sounds and short comments, so it does not require almost any effort to produce.

On the other side, storage is cheap and very easy to access and use. This is true not only for companies but also for private people. With the advent of the “Clouds” one of the most successful service is indeed storage, and a new way of making backups. Gone is the time when backups where done on tapes and it took a long time to make and, in case of necessity, to retrieve them. Today we typically have primary storages, even on multiple devices, for immediate use and a secondary storage where data is archived but it is still accessible directly even if more slowly. Dimensions of these archival storages go from Giga Bytes to Tera Bytes without much difference if not in an increase in costs. Now “BackUp” is actually remote on-line data archiving.

Apart from other security issues of this type of “BackUp”, which we are not going to go in detail

here, one major consequence is that there is not any need to select and sort data to keep, everything which is produced is kept since there is not space constraint to limit us. We always think to have a look to what is in the storage and delete something irrelevant, not useful, possibly harmful etc, but actually very few people ever do this.

So the status is that production of digital information is very easy and pervasive, and practically all data produced is permanently stored and archived.

Is this good?

To BackUp or not to BackUp

The life-cycle of digital information can be grossly summarized as follows:

1. digital data is produced
2. digital data is stored in the primary storage
3. digital data is archived for long (or "infinite") time.

Years ago each of the three steps had technical limitations:

1. it was not easy to produce digital data, in particular in large quantities
2. primary storage was limited and costly, and often to store some new data it was necessary to delete old data
3. backups where expensive, technically complex and took a long time to do; retrieval from backups was a long procedure and was not guaranteed to succeed; this usually lead to a careful choice of the data to backup and the rest of the data was discarded.

As of today all these three limitations have been removed, but have been removed also the filters which forced us to select and pick the information and data that we kept.

So, is this good?

The answer is No. But why?

There is another dimension in information, and this dimension is the meaning and value of the information. Looking at information from the point of view of its meaning and value, it is obvious that only "meaningful" and "valuable" information should be archived. But it is not so simple.

Some information can have a meaning and a value at one moment and can have a completely different meaning and value at another moment. Actually, in some cases the storing for long time of some information can have quite negative value, and become quite a large security exposure and reputational risk.

In a company the typical example is all the discussions, in digital form of course, done by internal email, chat etc., about some issues, or projects or personnels. Once all this information was exchanged in, sometimes heated, oral, face-to-face discussions in a close room, sometimes in a corridor, in a office, but they were only words which disappeared from the world as soon as the sound ceased, and remained in case only in the memory of those who were present. Now these discussions and the information in them, are in digital form, preserved for ever and, possibly, at everyone disposal.

So what could, can and has happened, is that part of this information, completely out of context, is one day retrieved and made public. The consequences can go from the hilarious, to the totally non interesting to the quite damaging to the reputation but also to the proprietary information of the company and its employees. This information can be considered, depending on the case, either as private correspondence, or proprietary data, or intellectual property and the (security) risk of its exposure can go from the non-existing to the maximum possible company damage.

Technology does not help us to manage this risk: this is mostly a human risk and technology actually is the main cause of it since it allows the misuse of the managing of the information.

Being mostly a human risk, it has to be approached by means of a Security Policy which instructs people on how to manage the information and what to do about it.

A security data retention and deletion policy

So how do we build a Security Policy for the correct management of information, retention and deletion, in a company? This is a question which brings us back many years, to the approach to information security of the XX century and in particular to military security, see for example the Orange Books [3] (or TCSEC) later evolved in the Common Criteria [4].

The first and most important point, is that information should be classified. It is not needed to impose a classification a la Secret Service, Spy Agency, Military or the like. What is important is to find criteria to distinguish different kind of information and data, and different ways of managing it.

For example, policies can be designed to allow only some type of information to be managed with some IT applications, and vice-versa that on some applications the data and information are managed according to very clear and well defined policies.

A simple approach to establish such a policy can be the following.

First it is established which kind of information must never have digital form. This is information that can be exchanged orally, or written on paper but should never appear in digital form. Even if it can look antiquate if not absurd, recently old style typewriters and their more modern descendants have been silently making a comeback, mostly for this reason [2].

It should be noted that the "never in digital form" ban by now applies also to telephone calls, in particular mobile calls or any kind of VoIP conversation. By now almost all telephone calls are digital communications, and the ban applies to them too.[2]

Actually this should not be seen as a bad issue, it should bring back the pleasure of discussions around a table with much more freedom of thought and expression which will never leave the room.

The second kind of information is digital *ephemeral* information. In this case the easiest way is to establish that information managed by some IT platforms have very short life and cannot be recovered after deletion. To this group typically belong application like chat, collaboration and social media, but also email and usual files on file-systems, which can live for minutes, days or just a few months and then be really erased.

Internal telephone calls usually are part of this group, whereas one cannot say a priori if external telephone calls are monitored, traced or taped. But this is generally true for all kinds of external communications, including chats, email and in particular social media information (which is never deleted). If one does not want to have surprises in the future, it is better to assume that all information exchanged externally will be stored forever somewhere out of our direct control.

Notice that for this kind of information we are establishing an (aggressive) *data deletion policy* since all data is erased after its short lifetime.

The third kind of information is any kind of digital data which is considered fine to keep for

2 We are not discussing here encrypted and secure telephone calls which are not common nor easy to implement.

sometime. An easy way to manage this can be to give to each user some but very limited archive space and leave to each person to decide which data to manually copy in this space so to preserve it. Data on this storage can be retained for example for maximum of 1 year and the combination of the limitation of space and of retention time should balance the risk of keeping any kind of information.

Obviously this is the most difficult point to implement of the entire policy. It is quite easy, after an initial success, to relax and allow the retention time and the space available to grow to practically unlimited values. If this is the case, the full purpose of the policy will be defeated. So it is very important that the limitation of this data retention area are enforced very strictly.

Finally, the last two kinds of digital information are the ones that must be kept for long time, for legal, regulatory or company decision. In any case, the company must decide very precisely which kind of information must be retained and exclude absolutely every either kind. This must be a short and precise list of information to manage in this way. Still there should be a retention policy which implies that all data must be erased at the end of its established retention period.

Actually there are two ways to archive this data, without special security measures and with special security measures.

In practice, some data can be extremely valuable to the company and its leakage can cause serious damages to the company. Since this data must be protected for a long period of time, it is necessary to implement a well designed encryption procedure for it. Typically data is digitally signed and encrypted with strong algorithms that offer protection for a long period of time. The encryption and signature keys should be stored in a very secure place different from the storage of the data. Periodically, for example every couple of years, it should be verified if the algorithms used to sign and encrypt the data still offer the expected security protection for the long period of time and in case the data is decrypted and signed and encrypted again with new algorithms.

Less important information that should be archived for long period of time, should still be protected but it is often not needed to implement such complex encryption procedures. Typically this data is archived off-line and access is carefully restricted and monitored. If it is archived off-site, as often happens also for disaster recovery procedures, it should be encrypted anyway, as tape backups are encrypted when stored off-line. The important point is that the recovery of this archived information requires careful authorization and that no direct access should be possible. This would prevent, even in case of IT intrusion, unauthorized access to the information.

Conclusions

Protection of information is more a procedural and human issue than a IT technical issue. It is more important that information is managed in the appropriate way according to its value and significance than to try to adopt global IT technical solution to protect it. The first step is not to produce in IT format sensitive information that can be managed in non-IT way. The second step, if information is produced in IT format, is to erase it in very short time, giving meaning and practice to its "retention" time. The third step is to archive in a well protected way and under strict access control, sensitive information that should be kept for long time.

As often in IT security, the problem is not only technological, by now we have many IT security instruments at our disposal, but we need to learn to use them in the appropriate way, even avoiding the use of IT when it is most suitable.

References

[1] For the "Sony Hack" see for example https://en.wikipedia.org/wiki/Sony_Pictures_Entertainment_hack and references within

[2] See for example "Unthinkable? Bring back typewriters" http://www.theguardian.com/commentisfree/2013/jul/12/unthinkable-bring-back-typewriters-editorial , and "Typewriters are back, and we have Edward Snowden to thank." http://www.washingtonpost.com/posteverything/wp/2014/11/12/typewriters-are-back-and-we-have-edward-snowden-to-thank/

[3] The Orange Book, DoDD 5200.28-STD, issued in 1983 and updated in 1985 by the National Computer Security Center (NCSC), an arm of the National Security Agency (NSA), http://csrc.nist.gov/publications/secpubs/rainbow/std001.txt

[4] "Common Criteria for Information Technology Security Evaluation" (abbreviated as "Common Criteria" or CC), ISO/IEC 15408, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50341

Andrea Pasquinucci (PhD CISA CISSP)