Preservation and Conservation (PAC) Programme
Frequently Asked Questions

Storing Digital Information for a Long Time

Prepared by PAC Poland

Q: Why do we store information?

A: Traditional libraries or archives keep books and documents for obvious reasons, namely, to access information stored in them – be it texts, graphics or sound. The objective of keeping digital information is precisely the same: one wants to be able to access it in the future.

Q: For how long do we want to store information?

A: The timing depends on the type of information. Generally, legal provisions determine time required for storing many types of documents, e.g. employee, financial or medical records.

Normally, they are to be preserved for several years to several decades. Cultural goods are usually kept for a long time in order to save them for future generations. Some documents are even meant to be kept perpetually (such as mortgage, land survey documents or vital records). In the case of digital information, similar time periods will most likely apply.

Q: What material is stored?

A: It is worthwhile noting that in order to store information recorded on traditional carriers we must keep those carriers. For instance, we keep documents on paper, photographs (negatives and positives, photographs on glass, paper and tape), audio discs, magnetic tape on spools and cassettes, featuring sound and/or images.

When digital information is regarded, in theory a similar approach could be taken, i.e. digital carriers with saved information, such as floppy discs and magnetic discs, CD-ROMs, semiconductor memories, just like traditional carriers – they can be kept on a shelf and read when particular information is needed. The experience of recent decades has shown, however, that information stored in such a way may be easily, and relatively quickly, lost.

Q: How long can information be stored using traditional carriers?

A: Old paper documents, manuscripts and books could be kept for several centuries or even longer without any special efforts. For this reason, one has come to expect that all documents could be stored in a similar way. However, new paper documents and other types of documents have proven to be less durable. In many cases, we have observed that carriers undergo a process of degradation, which entails the risk of losing the possibility to read information from documents.

One example is the so-called acid paper. It was manufactured starting from mid-19th century and used for printing newspapers and books – now however, publications on such a paper are falling apart. Degradation affects also traditional photographic materials and magnetic audio and visual tapes. In case of the former, the damage is caused by materials applied for their manufacturing initially. Some tapes after several years or several decades cannot be read anymore because the magnetic layer is detached form the underlay.

Apart from such radical form of degradation that renders reading impossible, gradual deterioration of quality of audio or film material can be observed, which takes the form of noise, greying photos or changed colours. Some of the processes may be slowed down by securing proper storage conditions, e.g. low temperature storage – this however increases the costs of storage.

Q: How can one secure information on traditional carriers?

A: One of the methods of protecting documents which may be subject to degradation is copying. The point is not to copy them physically but to create duplicates that will maintain the information contained in original documents. An additional advantage is when a copy takes up less space than the original and is as precise as possible.

These two objectives tend to be contrary to each other so in practice one needs to strike a certain compromise. Microfilms were a successful form of such copies, insofar as they were based on materials of much higher durability than traditional ones.

Digital copies however have been deemed to be more convenient even though the durability of digital recording on utilized carriers is limited. A big advantage of digital copies consists in the possibility to manipulate them in an easy manner, and in particular to make further copies, compare them with others etc.

Q: What digital information is stored?

A: At libraries and archives, one encounters various types of digital information. These include copies, such as photos or scans, of physical documents stored at cultural institutions. A copy may be created in order to secure an endangered document, and to make the information stored by it available to others (irrespective of whether a document is endangered or not).

The other type are digital objects that are not copies of stored documents as they do not have a physical prototype but were created as digital objects in the first place. They are actually referred to as “born digital” data. Such objects may be new books or journals. They may contain bibliographic descriptions or textual documents for archiving, such as emails. Also, they may include digital audio or video recordings or data, e.g. various types of measurements taken with proper devices.

Audio and video recordings may be commercial products manufactured on many copies; alternatively, they may also be unique. For instance, measurement data, due to their nature, normally are such creations. Losing an object representing the first type may be reversible, after all one may have another copy made of a physical object. However, if a unique digital object is lost, the loss may be irreversible.

Q: What carriers can be used to store digital information?

A: One should differentiate between storing collections of carriers that form resources of libraries and archives and may be kept just like books, for example on shelves in special containers to protect them from dust and saving data “on the fly” on computer systems.

Digital carriers stored as resources of libraries and archives include magnetic tapes, CD-ROMs, DVDs and their modern counterparts: recordable CDs, DVDs, Blu-rays. Moreover, this group includes typical memory drives, such as magnetic discs and solid-state drives (SSDs), as well as flash drives.

Widespread digital carriers for storing information on computer systems include semiconductor memories and magnetic discs or their arrays. Moreover, professional magnetic discs and tape drives are available on the market, including whole sets of such carriers handled by robots.

Q: Is digital information endangered mainly by carrier degradation?

A: When storing analogue information, the life cycle of such information is limited by the life cycle of a relevant carrier; it is also endangered by carrier degradation. Digital information, on the other hand, may be damaged at an earlier point. Therefore, one must take into account potential threats which digital information is subject to, as well as perspectives regarding individual types of carriers.

Obviously, the same digital information may be saved on various carriers, including those based on physical recording mechanisms. Such carriers take advantage of two physical conditions, i.e. magnetic material is magnetized in one of two directions, whereas a semiconductor capacitor is either charged or not. Such two conditions are marked with 0 or 1, i.e. two values of a binary digit, called a bit. Normally groups of bits are applied, for instance to save alphabet characters. Typically, eight-bit groups (called bytes) are used, as well as groups with 16, 32 or 64 bits and more. Such a division into multi-bit groups is of a conventional nature, and they are entered as a series of zeroes and ones.

Putting the physical aspect aside, i.e. whether magnetic, optic or other form of recording is applied, one should focus on digital information itself. It may be described on two levels.

At the first, technical one, one will only see bits, i.e. sequences of zeros and ones. At the other, in the same sequence, one may differentiate bit groups and information assigned to them. A bridge between these two is an agreement on how information is assigned to bit groups. Of course, there may be many potential methods and when reading, the same interpretation means must be applied at the point of recording to ensure correct reading. This is where formats, standards, and their conscious and knowledgeable application come in.

Q: Will stored information be subject to changes?

A: Any changes of stored information are undesirable. Still, one needs to be aware of the fact that independent of what physical mechanisms form the basis for a given type of memory, in every case the recording may become damaged.

Let's assume than a change affected one bit. The saved and the read sequence will not be the same anymore. Such information will be read differently. From the formal perspective, we will speak of a loss of original information in such cases. Moreover, we may overlook this fact at the point of reading.

Of course, one bit is just a tiny part against the backdrop of thousands or millions of bits used to save digital data. One may come to the conclusion that such an insignificant change should only lead to negligible distortion of information. This holds true in many cases for analogue data, e.g. a change in polarity of a small part of a magnetic data carrier if it's a sound recording would probably only give rise to a small crack when the recording is played. One could speak of information distortion in this case rather than loss thereof. Digital recording shows a different sensitivity to potential changes, whereby the effect depends on the format in which the information is saved. In certain parts, a change of one bit may lead to negligible consequences, but it may happen that consequences will be grave – the worst scenario is that a change of one bit may render the playing of a file impossible. This would mean that all the information is lost. Let's keep this at the back of our heads that sometimes we can't play a file because of such changes.

One may ask general questions with respect to such changes, putting aside the issue of carrier type. It is worthwhile having answers to such questions. Moreover, one should know how long such changes could occur on individual types of carriers and what reasons could cause them.

We have already indicated what consequences a change of one bit might have. Obviously if any single bit could change, then such a change could easily affect also a greater number of bits. And such a scenario should be also taken into account.

Q: How can one protect information?

A: First of all, we should have a damage detection mechanism in place. This can be achieved in a way similar to account numbering applied by banks, where a checksum in the beginning of a number will prevent certain errors – it makes it possible to verify whether an account number is correct. Similar checksums may be applied to data saving. E.g. by a parity bit. For each data part, bits will be calculated and a bit amounting to one will be added in the case of an odd number of bits and a zero when it is even. As a result of this procedure, the number of bits in a given part plus the additional one will be always even and should stay this way at the point of reading.

Some means of protection are more potent as they allow for correcting certain errors. This is done by adding excessive bits in order to detect errors and correct them. Consequently, one must save more data. Selection of a proper solution is normally made by a device or software producer. On its basis the manufacturer may claim a longer period of reliable operation of a device.

There are more complex information protection systems as well, e.g. based on additional protection drives. By the way, the easiest manner of protecting data is to save a mirror on the other drive out of a set of two. However, normally such solutions are used to protect ongoing work and not to archived data.

Q: How can data be protected against loss due to carrier malfunction?

A: The basic solution is to back up the data. However, this is just the first step as one needs to determine how and how frequently the original should be compared with the backup copy or copies. And what should be done when they differ. Moreover, one must provide secure storage conditions for carriers, adjusted to a given type of carrier. The optimum solution is when one stores copies in a distant location from originals, so that if a certain threat (such as theft, fire or earthquake) occurs at one location, the other will not be subject to it. Some feel the most secure practice is to store data in three distant places.

Even at home or at your office, when you use USB drives or CD-RWs for backing up your data, it is smart to develop certain procedures of handling them. In particular, one should define how frequently the data should be compared. Then it is worthwhile adhering to such procedures.

Q: What is the expected durability of digital data recorded on typical carriers?

A: As a durability measure of data saved on magnetic discs, one may refer to the guarantee period indicated by carrier manufacturers. In the case of regular hard disk drives, it is normally 3 years, whereas if they are of top quality, the period will be 5 years. This period applies to daily operated drives and not necessarily to drives that are stored for three or five years on a shelf. After all, a daily utilized drive may refresh the data, which is not possible in the case of those on a shelf. For magnetic recording, one may safely assume that the durability will be about two years. Anyway, in the case of valuable data, one should not expect to store them on magnetic discs. And if so, one would need to rewrite such data from time to time, using similar or different carriers. The same rule applies to storing CDs, DVDs, and Blu-rays.

Initially, recordable CD-RWs seemed to be a particularly durable and secure carrier. Optimists estimated that they would last for 20 to 30 years. However, real life has proven otherwise. Manufacturers introduced discs of variable quality on the market, some of those would not endure even one year.

In general terms, optical discs suffer from light exposure. Some manufacturers have offered discs whose reflective coating is made of gold to prevent oxidizing, claiming such CDs could endure up to 300 years; in the case of DVDs the claimed lifespan was 100 years. There were some drawbacks to them as well though, demonstrated by internal instructions of American archives not to use such discs for archiving purposes but only for data transfer.

The third popular carrier type are semiconductor memories. Typically, producers guarantee that their longevity is two, three or five years. In some cases, producers go so far as to give a lifetime warranty. Interestingly, this implies that such memories have no defects in material and workmanship. It does not mean that no reading errors will occur. Note that such declarations are based on tests regarding the number of errors that actually occurred at the point of testing. Surely, if a recorded semiconductor memory is stored for many years, data will be lost due to the disappearance of electrical charge in memory cells.

Another solution includes producing COM microfiches and microfilms from the digitised files. These are much longer lasting than the digitised files, and if needed, can themselves be easily digitised again

Q: What are basic conclusions regarding storing digital data?

A: Digital information carriers used at present will cause data loss after some time – and, importantly, this is not a lot of time.

Backing data up will reduce the danger of losing data and is necessary, but it will not prolong the lifespan of data recording. Users who wish to store digital information for a longer time must adopt an active approach to it. For instance, they need to refresh the recording or transfer data to new carriers at regular intervals.

Irrespective of producers attempting to develop long-lasting memories, one must act in a systematic manner in order to gain high certainty that the information we store will last for a long time and will be readable in the future.

Q: What is long-term storage?

A: A systemic approach to digital data storage and archiving has been developed after some incidents of irreversible loss of valuable data that were stored in a conventional way and not refreshed on time. In the face of such events, concepts and models have been developed in order to determine optimum operation of digital repositories, so that the data can be stored for a long time. Taking into account a perspective of long-term data storage, one considered aspects that were ignored in the past. Namely, we have to regard both actual changes and those which can be foreseen for the future that could prevent reading or interpreting such data.

These days, computer manufacturers keep changing their devices, sometimes forcing us to take up their new products. This happens for example when certain older technologies are no longer supported, like in the case of old types of cassettes for data archiving. Moreover, new formats for data recording are emerging. Take countless formats of graphic data as well as formats applied in text editors. This makes it difficult reading files saved in formats which are no longer widespread. One is faced with the necessity of converting such files into new formats. But this implies costs, especially if one has to take into account copyrights of such formats. Therefore, it is advisable to use open formats, but this in turn may require yet another conversion from a restricted to an open format. Another big question is whether data may be stored separately from metadata or is it safer to store them together.

Q: What is the difference between long-term data storage from data archiving?

A: More specifically, one should distinguish between data storage and data archiving. The former emphasizes the information itself, whereas the latter poses additional requirements, the aim of which is to give confidence as to the information which may be read after a long period of time.

Long-term storage describes a situation when a user intends to store information longer than the lifespan of contemporary technologies (of carriers, devices and formats), and longer than one generation. Such information will be read by persons with a different background than those who saved it.

In contrast, the following requirements should be fulfilled for long-term archiving:

- Information durability (this requirement is difficult to meet looking at the properties of carriers and changing technologies);
- Verifiability of proper storage;
- Information integrity (completeness and confidence that no modifications have been introduced);
- Authenticity (conformity with actual content with the declared one, e.g. as regards metadata);
- Availability (possibility of searching and finding desired resources);
- Interpretability (ensuring e.g. dictionaries and ontologies used at the point of creating metadata or bibliographic descriptions; more widely speaking this requirement refers to respecting standards);
- Confidentiality (guarantee that such data will be made available only to authorized persons or entities).

Q: Are there any standards applying to long-term storage and archiving?

A: To guarantee the possibility of correct interpretation of archived resources, it is necessary to conform with standards regarding the contents of such an archive, as regards data formats and metadata as well as with standards describing the structure of the archive and procedures in place. If there are no detailed standards in place, procedures should be documented.

One of the most widespread standards is the Open Archival Information System (OAIS), which lays down a reference model for digital archives. Also, there are various standards determining the structure of archive packages as well as of metadata.