



Best Practices on Standards, Policies and Quality Assurance in Digital Repositories for Long Term Preservation

Olivier Rouchon, Philippe Prat, Mathieu Cloirec

(CINES)

Abstract

During the past twenty years, the long-term preservation of digital information has only been a matter under consideration for a few scientific or patrimonial institutions. These have played a key role in the understanding of the subsequent risks and the definition of standards in this domain. The best practices rely on four technological risks which are now commonly agreed: the loss of the knowledge of the content, file format obsolescence, aging media causing data loss, sudden software or technology changes. They have been put in place in institutions dealing with text, images, sounds or video where quality assurance procedures have been developed to guarantee the integrity and accessibility of the data. The way this translates into raw, primary data produced by Tier-0 systems will be evaluated as part of this whitepaper.

1. Introduction

The data created, handled, processed, stored, exchanged and distributed in our society is mainly in digital form (figure 1). This form of representing data is incredibly powerful and the storage costs are becoming lower and lower; it is now possible to preserve them without any alteration whatsoever, and tools exist for creating complex documents and finding useful information in them.



Figure 1 – Binary representation of data

And yet behind all these huge advantages, lies a major risk; that of severe vulnerability to time, explicable and identified.

This vulnerability lies in the inevitability, if no preventive measures are taken, of one or more risks linked to the nature of the data itself, i.e. (figure 2):

- Lost knowledge of the content of digital objects: to be able to manage and recover data, details must exist to reference and locate them. Without this description, the ability to find the information depends on the memory of the person who created it.
- Impossibility of reading the format of the files containing the data: data in digital form is nothing more than a series of 0's and 1's (figure 1), and a standard market format is needed to be able to recover the informational content. There are however a great many of these standards, and they are constantly changing; some formats widely used several years ago have completely disappeared.
- Deterioration and ageing of storage media: a complex read system requiring hardware and software is essential for accessing digital data. This system may alter over time, due to physical or chemical effects, such as electromagnetic radiation, heat or even dust, making it impossible to read the data correctly.
- Disappearance of read hardware or software: the durability of a read system depends on that of the manufacturers or on their will, as it may be the aim of software publishers to maintain your dependency on them. Unavailability of just one component in the read system may definitively compromise data assets. Everyone knows the speed at which digital technology evolves and how fast the new eliminates the old. Here, we are confronted with the short shelf life of software and hardware, pure and simple disappearance of storage technologies, one after the other. Such disappearance means that the hardware

used by this technology is no longer manufactured and disappears from the market, resulting in an exponential rise in the maintenance costs of existing equipment.

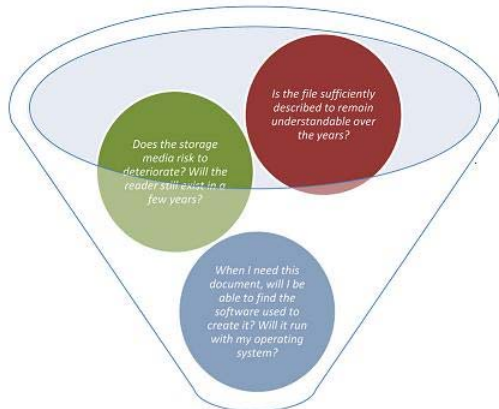


Figure 2 – The issue of preservation

Here, the time scale is a major factor given the issue at hand. Within a period of the order of ten years, the problem is (relatively) easy to deal with. Good quality and secure IT storage guarantees against accidental loss of the document. Technologies won't have changed so much that the document will have become irremediably unreadable.

Finally, the community of potential users of the document will most likely be scientifically and culturally similar to the one which created the document ten years earlier.

Within a period of thirty years or more however, none of this is a foregone conclusion, unless someone has thought of accompanying the document over time.

It is this period over the very long-term that lies at the heart of the long-term archiving challenge – i.e. the ability to mitigate the impact of the four above-mentioned risks on the day they occur, and that is exactly what quality assurance aims at, guaranteeing the intelligibility and accessibility of digital documents. The main methods to be implemented are now proven:

- The use of persistent metadata and identifiers to preserve knowledge of content,
- The choice of sustainable file formats to keep control and the ability to migrate to new formats (when conversion is the preservation strategy),
- Proactive management of the ageing of storage media for the ability to properly preserve the bit stream making up the files and migrating them to new supports,
- Permanent technological watch and anticipation of technological change.

The quality approach via best practices in this field can be seen from two aspects, technical and organisational.

2. The quality assurance approach to data

The technical quality approach to preserving digital documents covers all the procedures aiming at ensuring a high level of quality of the digital object itself. It can be divided into three levels.

2.1. The quality of the metadata

Metadata (meta: beside, beyond, including) are data which are used to preserve information describing digital objects; in this case we speak of preservation information (description of the information content, its origin or source, and its history) and content information (technical aspect of the information, its structure or format, and access rights). Several types of checks may be carried out, facilitated by the adoption of metadata standards: standardised metadata are described in a reference system that can easily be used to check the level of quality. The range of such sets of metadata is very wide, from the more generic for describing digital resources (i.e. Dublin Core, ISO 15836) to the more specific to a special field such as e-commerce (i.e. ebXML, ISO 15000) or geographic data (ISO 19115), via technical preservation metadata (PREMIS, METS) or administrative metadata for intellectual property and copyright (MPEG-21).

For intensive computing, a piece of data is an elementary representation of reality, and as such is not self-descriptive, and does not necessarily have an evident meaning. Here therefore, the aim of metadata is to add a sufficiently relevant descriptive level to enable it to be exploited in the best possible conditions.

Metadata are usable during the whole lifecycle of the data, by systems responsible for production, sharing, cataloguing, storage and access, and ultimately by the end-user.

A survey into the long-term archiving of scientific data conducted by the CINES in 2011 with 150 laboratories using intensive computing, shows that the communities where scientists work are very different, that the laboratories are highly specialized in their field of activity and that data description is not necessarily their first priority.

Most laboratories do not make use of standards concerning sets of metadata. They describe data mainly using references to text files, notes, publications, theses, web pages, source code, simulation parameters, or even just using a mnemonic file naming system.

Yet the main objective is simple: to describe data so that they can be exploited by all persons or systems to whom they are destined. To do so, measures at several levels can be envisaged:

- Organisational: find contacts, organise exchanges according to skills.
- Methodological: it is important to identify the recipients of the data (target community) and measure their ability to understand this data (by means of a knowledge base, for example). Then, a set of representation information needs to be implemented that will comprise a semantic link between the data and the community.
- I.T.: implement hardware and software infrastructure and protocols that will enable processing in the real sense of the metadata.
- Archivistic: specify and find pertinent standards and exchange formats in the field to enable the information to withstand the ravages of time.

Communicating information and collaborative work are essential in today's open world. Therefore, it is important that laboratories and scientific communities with a common issue, unite around these challenges so as to give rise to common resources and reference systems able to ensure readability and sufficient understanding of the data produced, with a guaranteed level of quality.

The main challenge seems to be the definition and standardization of intra-disciplinary metadata and their formalization. The example of the geography communities can be highlighted, since they developed the series of ISO 191XX standards for metadata usage, which is part of the INSPIRE European directive with the objective to establish an infrastructure for spatial information in Europe to support Community environmental policies, and policies or activities which may have an impact on the environment.

2.2. The quality of file formats

In order to be readable and convertible over time, files must strictly respect their format specifications; free tools (Jhove, DROID), and format-specific databases (PRONOM) which are authorities in the field of electronic archiving, are used to identify, validate and characterize files and ensure they are eligible for long-term preservation. In order to enable file quality to be verified, it must be in an identified and verifiable format. Then, file formats that are mastered must be preferred, i.e. formats whose specifications are published (i.e. PNG, ISO 15948) and standardized if possible (i.e. PDF, ISO 32000-1), and above all widely used by the community of information producers. With this in mind, the CINES digital preservation platform manages a limited list of file formats (around fifteen in total currently), and uses this open source software to conduct a whole series of quality controls on the files deposited before their long-term archiving.

High-performance computing is special in that it processes and produces significant volumes of data which must be sufficiently structured to ensure coherent representation of the whole and thus be understood in full by a person, an IT system, or both.

Whether it be for consumption, production or exploitation, the life cycle of the data requires the use of libraries, software or applications which are sometimes "in-house" and often determine the format of the data.

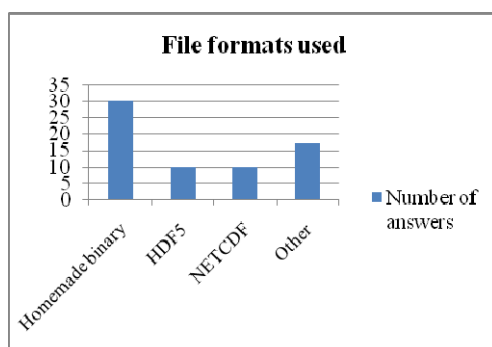


Figure 4 – Tape storage formats

A survey conducted by the CINES in 2011 with 150 French laboratories shows that most intensive computing projects have data in binary format at the output of software or calculation codes used (figure 4). ASCII and text files are used in one third of projects often in addition to calculation data. Data in HDF5 and NetCDF formats are also frequent.

Other formats are used more rarely but may be interesting to study (FITS, Grib, CGNS). However, one frequently encountered problem is data harmonization.

As the data cannot be used from one software to the next, they systematically need to be converted.

Faced with this issue, working groups have developed pivot formats, sufficiently specified to be used as standard. Here the objective is to design formats that are sufficiently generic to be understood and interoperable

within the same community while at the same time meeting the constraints of an issue often related to a discipline.

Of the most frequently used, we can highlight:

- HDF5 and NetCDF are open, copyright free and very generalist formats. They are self-descriptive in as far as the data and metadata are contained inside the file itself. They are designed to contain and handle matrices and/or grids.
- FITS is a format that is open, copyright free and designed for scientific images, enabling in-depth description of the image components using metadata contained in ASCII format in the headers. Each data block can thus be described using a keyword/value pair. While many keywords are reserved for the FITS format, the user is able to define his own.

As the technical specifications for these formats are published, tools to check the quality of the data files have been developed and can be used to validate their eligibility for long-term preservation, in the same way as the open source software mentioned previously.

Scientific data generally represent very specific phenomena with a certain complexity. Any description has its limits and if it has not been well thought-out, there may be a risk of loss of knowledge, if for example a person leaves. It is thus very important to describe data on at least two levels to limit this risk.

A syntactic description is used to know how the data are organised in the file (primitive types, size, position in a table, etc., are some examples). This information is generally included in the file headers and is mainly aimed at the IT systems used to exploit it.

A semantic description will provide information on the correspondence between the data and the meaning attributed to it. A value corresponds to a temperature or a pressure for example. These metadata may also be contained directly in a file or described in another file.

What can be done to maximize the chance of the data being exploitable by a third party in a context more or less removed from our work?

The aim is not to impose a standard format or a longwinded data description process to a laboratory that doesn't have the means to implement it, but to raise awareness and encourage producers to manage the risks and consequences of loss of data exploitability better.

It may be useful to pose a certain number of questions concerning the life cycle of one's data:

- For what reason were the data produced? (Was it with the view of sharing with a wide community, or only for the specific work of a laboratory?)
- Who are the recipients of the data, and do they have a knowledge base and tools able to exploit the data? Will laboratory members, for example, or all scientists working on the theme, be able to understand the binary file?
- How long will the data remain relevant and is it saved sufficiently in relation to its cost of production?

Depending on the answers, it may be useful, necessary or even imperative to launch migration of the in-house format without description to a standard format, and if this is not enough, to associate a set of metadata that is understandable by the recipients. Note here that our survey reveals that 40% of binary files do not, apparently, pose problems of migration to a standard format such as HDF5 or NetCDF.

However, a binary format with a good description of its content will remain easily exploitable by a third party and will probably last over time. Thus, BEST [1] is a tool proposed by the CNES to describe binary files, either syntactically with the EAST language, or semantically with the NASA internal standard (DEDSL) that is now internationally renowned.

2.3. Storage quality

Good preservation of the bit stream comprising the files containing the data to be preserved is crucial; it is the lowest level, the closest to the physical medium used to store the data. The first strategy consists of making multiple copies of the same document [1], if possible more than two. Determining the number of copies may be problematic, in particular from a financial point of view, as when the volume of documents to be preserved increases, the storage costs increase proportionally, potentially reducing the number of copies that can be made. It is also wise to use several storage technologies, ideally different, to avoid possible design or structural errors that may appear over time. Finally, a regular audit of all copies will detect any damage, and corrective measures can be taken if required.

Checking the integrity of files thus enables data corruption to be anticipated.

This can be done by the hardware, by CRC algorithm by the disk or network controllers, or by the software by running a checksum (comparing sampled digital fingerprints with the initial fingerprint using hash algorithms (MD5, SHA-256, etc.).

Several studies have been conducted into the reliability of storage media, their age resistance and silent corruption of the data, of which can be mentioned studies from the University of Carnegie Mellon [3], Google [4] and the University of Wisconsin-Madison [5] on hard disks.

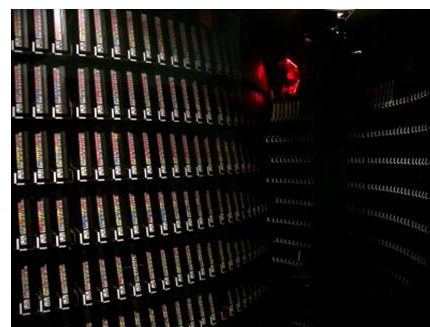


Figure 5 – Tape storage

The results are enlightening and an incentive to the greatest vigilance towards online disks.

For intensive computing, data volumes produced are such that the use of disk technology is far more costly than the use of libraries of tapes. Some studies (including those conducted as part of the WP7.6C for the PRACE-1IP project [6]) reveal that above a few hundred Terabytes to be stored and preserved, magnetic tapes are much more cost-effective and safe than hard disks. In any case, the same integrity check mechanisms apply to this technology and should be implemented.

Technical quality is therefore unavoidable, but is only relevant when the preservation process as a whole is organized with the same requirement for excellence.

3. Risk management as a tool for managing preservation

Project management based on risks originated in industry and is now widespread in the management and service sectors.

As preservation of digital documents is essentially a preventive measure, something must be done before the damage occurs. This is a set of compromises both between the immediate needs of data producers and the long-term needs of users, and between the needs and resources implemented.

The preservation of digital documents is a project like any other, and like any other activity, generates risks. The aim here is not to eliminate the risks, but to determine an acceptable level of risk. This is a well-defined and proven method.

To start with, the context of application needs to be defined, and the risk management objectives fixed; then the various phases are as follows (figure 6):

- Identification and classification of the risks. A risk is an event which, when it occurs, may prevent the successful completion of a service mission; it is defined by the combination of a vulnerability and a threat;
- Risk assessment, by analysing probability and the impact of each risk over time, and the combination of these two factors; there is a third assessment criteria which may be taken into consideration: the degree of appearance over time, assessed according to the degree of imminence of the risk over time. The risk priority index (RPI) can thus be calculated on two dimensions (impact x probability), as is done at the CINES, or on three dimensions (impact x probability x degree of appearance over time) as often evoked in presentations by the French national library on the management of risks applied to digital preservation, or those of the FMEA approach, more suited to the industrial context;

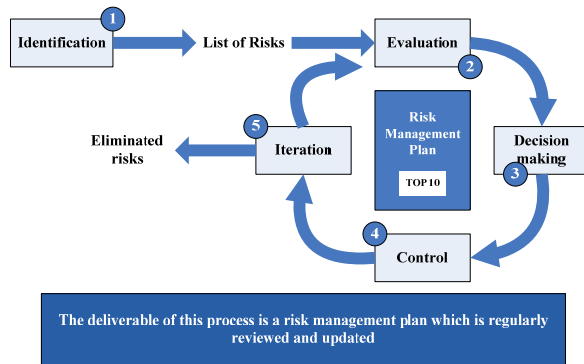


Figure 6 - The risk management process

- Decision-making: the priority risks need to be identified, according to the level from which they become acceptable, the way of dealing with them and the action plan; to do so, it may be necessary to compare them to the goals and activities of the establishment, to the costs and benefits of mitigation operations or to the legal obligations.
- Risk control, by implementing the measures required to reduce the level of risks, either to reduce their probability, or to reduce their impact.

This is an iterative process, and is only effective if it is regularly reassessed, as some residual risks may be eliminated, whereas others appear during the operation of the system.

4. Certification of a long-term archiving system

Certification is the culmination of consolidating an organisation and/or service. It signals recognition of quality and professionalism, and is therefore a means of instilling trust with communities of users, and may also be a way of leveraging budgets from governing bodies. Several types of certification can be envisaged:

- Generalist certifications, such as ISO 27000 (IT security), ISO 9000 (quality), CMMI (engineering), ITIL (services), etc.;
- Specific certification for durable archiving, including: DSA (best practices, figure 7), DRAMBORA (risk management), TRAC (list of criteria) and certifications currently being drawn up: AFNOR (Z42-013) with the SIAF, ISO 16363 (European Audit Framework) with the CCSDS.



Figure 7 – The Data Seal of Approval

Certification is a weighty project, both in the human investment that is required to manage the project or the changes required, and in the financial investment, as regular external audits needs to be planned. As a result, it is important to identify the type of certification that will have the greatest impact both on the community of users and on the governing bodies.

5. Conclusion

The ease of preserving digital objects over time depends largely on their quality; however quality should not be limited to the technical aspect of the digital objects alone, as the quality of the preservation process applied to them is just as crucial.

Awareness of the need to implement standards, and respect for certain best practices, in particular regarding a quality approach to long-term preservation, is not neutral. It represents a considerable and immediate investment, the effects of which are only perceptible in the long-term; moreover, such an initiative, and its culmination in certification, requires the commitment of everyone involved in the preservation process.

However, quality indicators and measurements are not yet clearly defined, with little hindsight available to the institutions implementing it to measure the positive effects. So, see you in thirty years?

6. References

- [1] Best framework - <http://logiciels.cnes.fr/BEST/FR/best.htm>
- [2] D. S. H. Rosenthal, "Bit Preservation: A Solved Problem? ", Stanford University, 2008
- [3] B. Schroeder, G. A. Gibson, "Disk failures in the real world: What does an MTTf of 1,000,000 hours mean to you? ", Carnegie Mellon University – Computer Science Department, 2007
- [4] E. Pinheiro, W. D. Weber, L. A. Barroso, "Failure Trends in a Large Disk Drive Population", Google Inc, 2007

- [5] L. N. Bairavasundaram, G. R. Goodson, B. Schroeder, A. C. Arpaci-Dusseau, R. H. Arpaci-Dusseau, “An Analysis of Data Corruption in the Storage Stack”, University of Wisconsin-Madison, 2008
- [6] F. Marceteau, “Media and technology appraisal for long term preservation“, PRACE-1IP whitepaper, 2011

7. Acknowledgements

This work was financially supported by the PRACE project funded in part by the EUs 7th Framework Programme (FP7/2007-2013) under grant agreement no. RI-211528 and FP7-261557.