IOP Conf. Series: Materials Science and Engineering **919** (2020) 052003 doi:10.1088/1757-899X/919/5/052003

Digital media inventory algorithm for long-term digital keeping problem

Alexander V Solovyev

Institute for Systems Analysis Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia

E-mail: soloviev@isa.ru

Abstract. The previously considered problem of long-term digital keeping involves a multi-faceted integrated approach to its solution. This article is devoted to the algorithmic solution to the problems of reliability and technological aging of digital storage media, as well as changes in the hardware and software storage environment that occur over the long-term digital keeping. The article provides a brief overview of digital storage media with an estimate of the terms of reliable guaranteed storage of data on them. The problem of changing the hardware-software environment for storing digital data, the cause of which is the limited life of the software and hardware of information systems, is considered. As an algorithmic solution to the problems presented for the problem of long-term digital keeping, an algorithm for inventorying digital media is proposed. Its main aspects and rules of practical application are considered. The connection with the previously proposed algorithm for controlling the authenticity of digital data is indicated. The article concludes that the proposed algorithm, tested by practice, will solve the problem posed in the article. In conclusion, the development prospects of the proposed algorithm are considered.

1. Introduction

In the articles of the author [1, 2], the problem of the long-term keeping of digital data (DD) is considered in detail. At the same time, it has been proved in detail that the problem is multidimensional and can only be solved in a comprehensive way by creating long-term keeping technology, which is a combination of methods and algorithms for solving individual particular problems. In [2], the problem of controlling the authenticity of digital data for the problem of long-term keeping is considered, an algorithmic solution to this particular problem is proposed. In this article, we consider the problems of reliability and technological aging of digital storage media, as well as changes in the hardware and software environment for storing digital data that occur over a long storage period (decades).

One of the most important problems in the long-term keeping of DD is the problem of failure of storage media (disks, tapes, optical media, SSD (Solid State Disk), etc.). No manufacturer of such equipment guarantees its safety for decades (especially centuries), and, therefore, the problem arises of the timely diagnosis of digital media and their timely transfer to other media.

This article is devoted to solving the problem of control of digital storage media during long-term keeping of DD. The paper proposes an algorithm for inventory of media, which can be used to solve the problem in the framework of software for hardware-software digital data storage systems.

2. Digital media reliability problem review

The warranty period for reliable storage of most hard drives is 5 years. Manufacturers of write-once optical discs (media like WORM - write once read many) initially called dates of 50-100 years, but then they were significantly reduced (in addition, they need ideal storage conditions) to a maximum of 20-25 years, after which the data should be overwritten. Based on the author's experience in creating electronic archives (EA) using DVD-Rs of leading manufacturers, the author may argue that in practice the shelf life of DVD-Rs is even lower, checking and dubbing should be done at least once every 5 years.

Even for ultra-dense recordings specifically designed for EA drives based on UDO technology (Ultra Density Optical, developed by Plasmon) [3], the possibility of their operation for many decades has not been confirmed. UDO drives are used, for example, for storing medical images, medical documents, medical records of patients. Moreover, the guaranteed shelf life does not exceed 5 years. UDO is a 5.25 ”cartridge with an optical disk inside. The disk capacity at the moment is from 60 GB to 120 GB. For recording, both a red laser (650nm) and blue-violet (405nm) can be used, and in the second case, the maximum disk capacity can reach 500 GB. An optical disk is not susceptible to demagnetization, like magnetic media.

The magnetic tapes used for backup are extremely unstable to external influences of carriers. For example, the problem of magnetic tape degradation is widely known. For more or less reliable storage, rewinding once every six months and careful protection of the tape from demagnetization are also required (see, for example, [4, 5]).

The use of solid state drives (SSDs, flash cards, etc.) is also not yet reliable. These drives have a limit on the number of rewriting cycles (3000-10000), increased wear due to this, the high cost of a gigabyte of information compared to hard drives and optical disks, and a low amount of data storage [6]. They are trying to overcome the problem of increased wear using FRAM (Ferroelectric Random Access Memory) technology for which the number of rewriting cycles is estimated to be of the order of 10^{14} [7]. However, even these media do not allow storing large amounts of data, but are notable for their high cost. Guaranteed data storage time on SSD and FRAM is estimated at 10 years.

Thus, industrial means of storing digital data at the moment cannot reach the maximum shelf life of information, such as on paper or in the form of microfilms: up to 500 years under ideal storage conditions.

In addition to the problem of storing digital data on a specific digital medium, there is also the problem of technological aging. Therefore, with a fairly high probability, after 100 years it will be impossible to read data from modern digital media due to the lack of devices for reading them in the future, even if the information is somehow stored on them.

An analysis of modern technologies (see, for example, [8, 9]) gives the impression that manufacturers are not very interested in the long-term existence of certain carriers, the average life of technologies from the moment of their appearance to their almost complete disappearance from the market is estimated at 10- 15 years (magnetic tapes, floppy disks, CD-R, DVD-R, etc.). Then new technologies supplant older ones, and it will be economically unprofitable for manufacturers to support outdated technologies.

3. The problem of updating the hardware-software environment for keeping digital data

In addition to the actual wear and tear of physical storage media, there is a closely related problem of updating the hardware-software environment of long-term keeping of DD) [10, 11].

The situation here resembles movement in a vicious circle: old operating systems (OS) and other system software are removed from support, the new one requires more processor performance, memory, etc., i.e. hardware update required. Old OS and other software cannot function on new hardware, etc.

For example, Microsoft's widespread Windows OS adheres to the following OS development strategy: for the first 5 years, the OS version is fully supported (updates are issued that can be installed by any registered Windows users). Over the next 5 years, extended support can be provided for a specific customer. Further, the OS is removed from support and its performance on new hardware is not guaranteed.

Briefly, the process associated with updating the hardware and software environment can be illustrated in table 1. The table shows the problems that arise over a 20-year shelf life, while the shelf life, for example, of so-called personnel documents is 50 years.

Table 1. Problems updating the hardware-software environment for keeping digital data.

Shelf life	Software Environment Issues	Hardware Issues
3 years	1) expiration of electronic signature (ES) certificates	
5 years	1) the deadline for the validity of ES certificates under Federal law of the Russian Federation 6 april 2011. №63-FZ «On the electronic signature» 2) significant improvements to EA software are needed	1) failure of storage media (CD, DVD, HDD, flash drives, etc.)
10 years	1) removed from OS support 2) file viewers of various formats become obsolete 3) proprietary data formats are no longer supported	1) SSD and FRAM service life 2) problems with the modernization of technology
20 years	1) significant limitations when using an outdated OS and outdated viewers 2) OS can only be used in a virtual environment 3) possible problems even with open data formats 4) changes in cryptographic standards are possible, as a result of the inability to verify ES	1) hardware storage life

Thus, digital data keeping must be ensured in an inevitably changing storage environment. At the same time, make this environment either trusted or use some means of ensuring the integrity of digital data and its metadata.

Both in the case of media aging and in updating the hardware-software environment, the need arises for the correct migration of digital data. The complexity of organizing data migration is that there is a high probability of data loss due to negligence or malicious intent.

The migration problem concerns not only the transfer of the digital data itself, but also its metadata (see, for example, GOST R ISO 23081-1-2008 “Document management processes. Metadata for documents”). If metadata: indexes, metadata, classifiers, categories, links with other documents, etc. cannot be correctly transferred, then, in fact, the migration of digital data will result in the re-creation of EA in a new hardware and software environment with the re-creation of all metadata, which is a task almost impossible. Incorrect migration can lead to loss of semantics of digital data.

4. Digital media inventory algorithm

As you can see from the review of problems, all the types of storage media currently available are not reliable enough to store data for decades, and even more so for centuries. Moreover, due to the process of technological aging, after a few decades there will be no devices that can read current storage media.

It can be argued that the solution to the problem of “aging” lies, firstly, in the redundancy of information storage, and secondly, in the regular verification and transfer of information to new data carriers.

Redundancy of data storage should be ensured by both storing EA data directly in the database on the hard disk and storing copies of EA data on external media. Such a copy can be either a backup copy of a database or copies of data that has been squeezed out to external media.

In all cases, for the EA database, it is necessary to organize regular backup of the database to external media.

A copy of the data can be either external media with a backup copy of the database, or a combination of external media with a backup copy of the database and external media with data. At the same time, at least two copies of EA data should be created, and they should be stored in different rooms, and ideally in different buildings, remote from each other. If you use CDs as a backup, it is recommended that you create at least three copies of the data.

Additionally, a solution resistant to external influences can be implemented (a mirror, or, for especially valuable documents, a backup data processing center (DPC)), i.e. storage of an exact copy (copies) of documents. This means that it is necessary to implement decentralized storage of copies of data with different access credentials for the operational and administrative personnel of the EA.

As a solution to the problems described in this article, the author of the study proposed an algorithm for inventorying digital storage media.

Regular verification and transfer (in terms of GOST R ISO 15489-1-2007 – conversion or migration) of digital data to new media should provide protection against failures and physical degradation of digital storage media. We call this procedure an inventory of media.

This procedure should include checking the integrity of the DD on the medium, evaluating the remaining storage time of the data on the medium, and, if necessary, transferring the DD to a new medium.

In the event of a violation of the integrity of data on the medium during verification, a new copy of the data is created from other copies of this information. Periods of verification of data carriers are selected based on the type of information carriers, but in any case, the period of storage of data on an unchangeable medium should not exceed three years, i.e. every three years, each storage medium must be checked and replaced if necessary.

The process of transferring information should provide for the possibility of merging data from different media, this condition appears due to the constant increase in the volume of all types of data carriers. Therefore, next-generation media, as more capacious, can contain information from several "old" media.

The proposed algorithm is presented in general form in figure 1. Only in case of its implementation within the framework of a specific system of long-term keeping of DD will it be possible to talk about the safety of DD during the “aging” of media.

All information on the progress and result of the inventory should be displayed in the system journal of the inventory of EA media. Also, information on the verification of authenticity before and after data transfer should be reflected in the system log of authentication inventory [2].

The problem of updating the hardware and software environment for storing digital data, as well as the problem of “aging” of media, necessitates the organization of periodic data migration.

It can be argued that data migration should be an integral part of the technology for creating any information system designed for the long-term keeping of DD.

In order to properly organize the process of data migration, it is necessary to answer the following question. What should be subjected to migration: are only the documents themselves from the EA database or other metadata, classifiers, indexes, audit logs, etc. related to them?

The answer to the question “what to transfer” should be a mathematical model of digital data during long-term keeping. All components of the data and their purpose should be spelled out in the model.

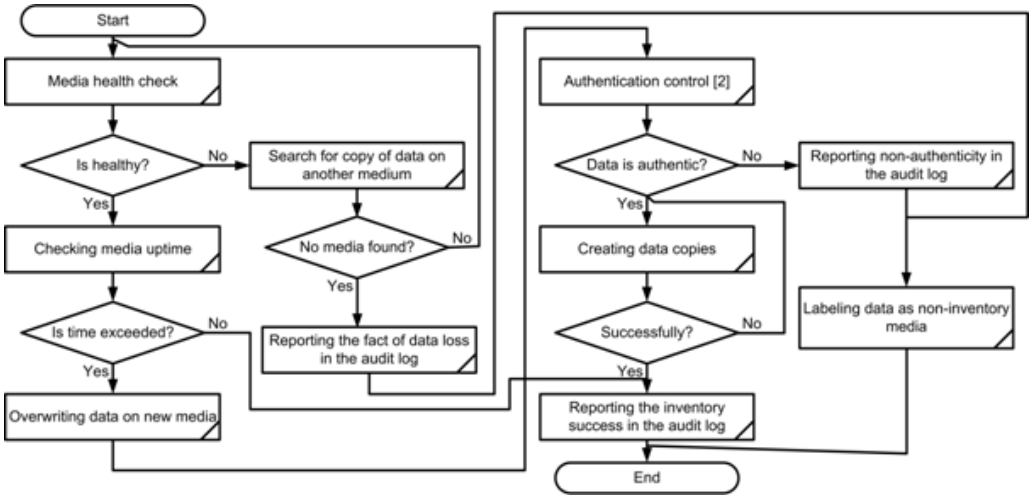

Figure 1. Digital Media Inventory Algorithm.

5. Implementation of algorithm

The algorithm proposed in the work was brought to software implementation. When creating a large geographically distributed information system - the electronic archive for the Pension Fund of the Russian Federation algorithm, a software module was developed for the inventory of various types of media, including removable ones. The created information system has been operating since 2004 in 80 regions of the Russian Federation and currently provides storage of more than 50% of the accounting documents of the Pension Fund of the Russian Federation.

The results of the application of the algorithm in the framework of a large information system helped to solve one of the important tasks of ensuring the long-term keeping of DD – ensuring the physical safety of data. During the operation of the electronic archive, 2 migrations were carried out without data loss. Thus, the digital data has been completely saved.

The algorithm was also tested as part of the creation of a number of electronic document archives and digital data repositories.

6. Conclusion

To solve the problems posed in this study of reliable long-term keeping of DD on modern media, an inventory algorithm for digital storage media is proposed.

The proposed algorithm was implemented in the form of software modules and was tested by practice in the development of a number of information systems of electronic archives, in which long-term keeping of DD was assumed.

In further studies, it is planned to significantly increase the productivity of software implementations of the proposed algorithm, to improve the algorithm itself in terms of ease of implementation. It is also planned to develop a mathematical model of digital data for long-term keeping of DD.

In the future, it is planned to finalize the technology created by the author for organizing the long-term keeping of DD, including an improved algorithm in it. It is planned to use the developed technology for enterprise-wide information systems (ERP).

References

[1] Solovyev A V 2020 Long-term storage technology of digital documents *Lecture Notes in Electrical Engineering* **641** 901-11 DOI:10.1007/978-3-030-39225-3_97

[2] Solovyev A V 2020 Authentication control algorithm for long-term keeping of digital data *IOP Conference Series: Materials Science and Engineering (MSE)* **862(5)** 052080 DOI: 10.1088/1757-899X/862/5/052080

[3] Arhivniye opticheskiye nakopiteli Plasmon UDO2 G-serii. ELAR [Archive optical drives

Plasmon UDO2 G-series. ELAR] Access mode: http://www.plasmon.ru/udo2_g.pdf (12/06/2020)
[4] Rivers M 2009 «Baking» Magnetic Tape To Overcome The «Sticky-Shed» Syndrome [Electronic resource] Access mode: http://www.audio-restoration.com/baking.php
[5] Gracy K 2007 Magnetic Tape Preservation: An introduction *The Association of Moving Image Archivists* Access mode: https://web.archive.org/web/20071011092011/http://www.amianet.org/publication/resources/guidelines/videofacts/intro.html
[6] Orlov S 2010 Nastupleniye SSD *Jurnal setevih reshenii/LAN* **11** Access mode: http://www.osp.ru/lan/2010/11/13005552/
[7] Rzehak V 2012 Osobennosti primeneiya FRAM mikrokontrollerov Texas Instruments [Texas Instruments FRAM microcontroller application features] *RADIOLOZMAN* Access mode: http://www.rlocman.ru/review/article.html?di=113273
[8] Korepanov I 2008 Kak sohranit' arhiv na desyatiletiya? [How to keep archive for decades?] «*Jurnal setevih reshenii/LAN*» [*Network decision journal*] **03** Access mode: http://www.osp.ru/lan/2008/03/4899898/
[9] Tikhonov V 2006 Arhivnoye hranenie elektonnih dokumentov: problemi i resheniya [Archival storage of electronic documents: problems and solutions] *Deloproizvodstvo i documentooborot na predpriyatii* [*Paperwork and document management at the enterprise*] Access mode: http://www.delo-press.ru/articles.php?n=5150
[10] Afanasyeva L P 2005 Automated Archive Technologies *Federal Agency for Education. State Educational Institution of Higher Professional Education Russian State University for the Humanities* 114
[11] Miller J 2012 NARA to suspend development of ERA starting in 2012 Access mode: FederalNewsRadio.com http://www.federalnewsradio.com/?sid=2204570&nid=35