Improving the Quality of Backup Process for Publishing Houses and Printing Houses

To cite this article: N E Proskuriakov et al 2018 J. Phys.: Conf. Ser. 998 012026

View the article online for updates and enhancements.

Related content
- Backup key generation model for one-time password security protocol
  N Jeyanthi and Sourav Kundu
- Files synchronization from a large number of insertions and deletions
  Vijayan Ellappan and Savera Kumari
- The "Nauka" Publishing House is Preparing for Publication a Collection of Articles "Sorbtsiya i Khromatografiya" (Sorption and Chromatography) 20 Sheets, Price 3 Roubles
  A M Sladkov
Improving the Quality of Backup Process for Publishing Houses and Printing Houses

N E Proskuriakov 1, B S Yakovlev 1, V V Pries 1
1 Tula State University, 92, Lenin av., Tula, 300012, Russia

E-mail: vippne@mail.ru

Abstract. The analysis of main types for data threats, used by print media, and their influence on the vitality and security of information is made. The influence of the programs settings for preparing archive files, the types of file managers on the backup process is analysed. We proposed a simple and economical version of the practical implementation of the backup process consisting of 4 components: the command line interpreter, the 7z archiver, the Robocopy utility, and network storage. We recommend that the best option would be to create backup copies, consisting of three local copies of data and two network copies.

Key words. Backup, data protection, cloud computing, network storage.

1. Introduction
Digital technologies are actively used in all branches of human activity. In polygraphy, they are especially important at the stages of preparation of layouts, communication with customers, organization of printing on demand, etc. These technologies are well developed and proven effective. Without them it is difficult to imagine a modern cycle of production of books and other printed products.

During the work of the publishing house, the customer base is accumulating, their orders are executed, which are inexpedient to delete, because there is a possibility of their repeated order. In the course of working on projects, a creative search of ideas takes place, which very often turns into a collection of a clipboard that can nudge certain ideas. Collections of fonts, collages, backgrounds, sets of brushes and styles for various programs - is a very important part of the results of the work of the publishing house, their employees, to some extent, the history and portfolio of the organization. Also worth mentioning is the software, their versions, device drivers, setup color profiles, etc. It is also an integral part of the collection of technological developments.

In addition, you need to take into account that the software version is very important component of a successful re-release, re-release of products. For example, during publishing books in the 90s Lexicon, Photoshop 4.5 editors were used, and it is not necessary that modern versions of the same programs will be able to open older files. There may be conflicts of formats, settings, etc. Because of this, previously prepared projects may not be available for replication.

Moreover, the old software distributions cannot be installed on a modern operating system without additional and not reliable software. So Windows 98 or 95 need to install Photoshop 4.5, and this image of Windows should be available.
In addition, in the current realities the virus attacks and competitors can harm the publishing houses, by stealing, destroying, corrupting the collection, practically stopping their work. According to the data of [1], in the current realities the ways of infection have become more diverse due to the development of social networks, systems of communication between users, the affordability of creating websites, mailing messages. Since data exchange between authors and enterprises of printing provides for a large data flow, infection cannot be excluded due to the employee's inattention. This issue is particularly acute escalated now, when there were the flashes of cyber-attacks with virus-encryptors for Internet users. Many large private and public firms, services suffered. Administrators of computer networks in the small and medium-sized organizations usually do data recovery after such attacks. However, the availability of special service for the information protection does not guarantee the success of data recovery. So, in the case of virus-encryption viruses, encryption takes place over a long enough password to be considered not to be opened. Besides, the password is changing in the course of infection; therefore, even in case of successful password guessing, it is suitable only for one specific file. Decoding in this case can last for months and years, and it will be too long for the company, and practically everyone, in such cases, consider then data is lost forever. Under the influence of such circumstances, the creation of the archive and its protection somehow still rises before the majority of printing enterprises and publishing houses, which is an actual task in modern production.

2. Formulation of Problem
To date, the majority of various firms’ and companies’ experts believe that for more high quality and reliable preservation of information it is necessary to copy the files on the devices of different connections and locations. Most often, the right technique is to store the files on the logical drives (hard disk different from the original), on removable, external hard drives and network storages. In this case, the information is systematically distributing in three different data stores what there is a very positive effect on the reliability of its storage. It is also possible to disconnect the external hard drive and sync with the cloud service, which will greatly improve the stability of the archive to virus attacks and malfunctioning. However, these statements are little supported by facts; the speed of the process has not been analyzed, which has become a very important factor recently.

In order to find out, it is necessary to understand which of the proposed or existing solutions will be most suitable for printing enterprises. Therefore, we have tested the three most popular software products and the rate of their work for copying files in two different directories of a personal computer (PC). The question was investigated - in what form backups should be created: as is or with the use of the archive; as multi-volume archive or not; what should the format of archive and its compression ratio be [1]. Beyond that, we also considered the question with network data repositories and their role in modern status of backup.

To identify the type of archive and the convenience of its use, we conducted a preliminary series of experiments, during which it was revealed that not all archivers could cope with the task assigned to them. Therefore, the archivers WinRAR and ZIP are not able to pack all files because of restrictions for file name length. Furthermore, the ZIP cannot be multi-volume archive and does not support large volume archives, such as our test directory of 31 GB 548 MB.

Consequently, we excluded archivers WinRAR and ZIP from further studies as unreliable because they cannot guarantee an error free backup with any degree of nesting objects. As a result, we left with only one variant of the archiver - 7z. This is a basic archive of UNIX systems, including Linux. Its main advantage was the fact that it is able to back up any depth of nesting information. It also supports the creation of multi-volume archives and the command line, being practically a utility [2]. In addition, its compression ratio is equal to results of WinRAR [3].

Next, we need to figure out which method of backup would be most beneficial. For this, we need to determine the degree of compression and volume of the archive file.
The compression ratio directly bound to the types of files, which take part in process and sometimes may be the situation when the archive with the maximum compression will decrease insignificantly. Therefore, it makes sense to choose the speed of backup process as the main benefit measure. The compression rate depends on the degree of compression of the file, for this reason, it is expedient to choose the no compression variant.

As for the division of the archive into volumes, this is not quite an explored area for making backup copies. Almost all well-known software solutions boil down to the fact that each project file is a separate ZIP archive. It is precisely unknown, how reasonably this provision is. Based on practice, we assume that this approach is wrong, because it does not reduce the number of files. Therefore, we believe that one of the main tasks of the backup operation is to create a backup reducing the number of files and to pack materials, which are heterogeneous in type, number and volume into blocks that strictly divide into a predetermined size. According to our observations, this clearly reduces the time of copying process. However, it is necessary to study in depth that we try to carry on.

In addition, the division of archives into volumes is a necessity in case of backup on the cloud services. The fact that almost all of the file storages limit the speed of file transfer because of the huge number of users, which constantly transfer data to the server. The communication channel cannot withstand such a stream, therefore the majority of servers limit the download speed to itself, while leaving the upload speed of files is almost the maximum. Moreover, this objective reality exists in the Internet. However, to verify this assertion it is necessary to conduct the study.

We investigated three types of software to synchronize with network storage: Google Drive, Yandex Disk and DropBox.

Synchronizing your directory with the network storage always comes at the expense of the software installed on the machine where the backup copy is. The task of software is to monitor the status of files in the controlled directory, and in case of their changes instantly to send files to the server. The server produces the creating of actual file versions, their storage and other service functions, such as cloud or network storage, but not the user software installed on PC.

Most of these programs have the standard settings: choice of directory for synchronizing, login and password for the connection. In addition, they can pause their work and embed into the system, reflecting in it as additional directories.

Network storages are comfortably to use and very well fit into the backup system. In such services there is version control, and even if something is infected on your PC and this will be reflected in network storages, then this backup can be rolled back using the version control feature.

Unfortunately, Yandex Disk does not support control of version. Moreover, it is not entirely clear how the history of file works. Therefore, we cannot guarantee the stability of backup in case of constantly changing files, used on Yandex Disk technologies. For example, we use Yandex Disk as additional hard disk drive.

In most cases, the network storages selected as one of the alternatives for the storage of backups, and usually the main reasons are:
1. The server can store a maximum of 5 versions of files that have been uploaded. This means that the owner of the archive can restore to an adequate version of the project at any time. Most often, this need arises from the malfunction of hardware, software, operating system, or when virus infected the OS.
2. Fulfill the rule: to store data on various media unrelated to the primary place of storage. I.e. you cannot store everything in one place; you need alternative data banks so that you can recover corrupted or lost data.

Based on this, we conducted a study of network storage for the best result. Various volumes of archives were created along the length of the archive and it was revealed which network storage was preferable and at what size of the archive volume.
3. Results of Experiments

We chose Google Drive, Yandex Disk and DropBox as network storages. We had to cut the archive into volumes of 100, 300 and 500 MB. In this case, we only considered the full volume size of the archive, because only they had a stable size, so we excluded from the studies the last volume of the archive. When splitting archive (volume of 31 GB 548 MB) into volumes at 100 MB, number of them in fact will be 300 pieces, 300 MB - 100 and 500 MB - 60.

The results of the calculations we give in Table 1 and in Figure 1. According to Figure 1, we can determine the speed of loading a test directory for each of the network storages. Based on the analysis of Figure 1 we can conclude that each software for file network storage creates different conditions for the processing of files depending on their volume. So both Google Drive and Yandex Disk increase the speed of receiving files for file sizes from 100 to 300 MB, but nearby to 500 MB - decrease. Most likely this is due to attempts by companies to make file downloads more stable and reduce the number of failures. DropBox uses a different strategy, simply increasing the speed of downloading files depending on their size.

Table 1. Estimated download time (min) of the whole archive on the network storage depending on size of volume (The total volume – 31.55 GB).

| Splitting into volumes [MB] | Network storage |
|---------------------------|-----------------|
|                           | Google Drive    | Yandex Disk   | DropBox |
| 100                       | 111             | 219           | 231     |
| 300                       | 80              | 165           | 188     |
| 500                       | 82.2            | 174           | 154.8   |

![Network storages](image1)

![Copying of directories in not archived/archived forms](image2)

The data from the studies allow us to conclude that for networked storages it is more expedient to use the division of archives into 300 MB volumes. In the next step, we have begun to identify the factors influencing on the process of copying files into logical drives. We presented research results in Table 2-3 and in Figure 2.
Table 2. Time of copying (min) the test directory with a volume of 31.55 GB in non-archived form.

| Media type information, that receives copy | File Manager | Total Commander 8.01 | Windows Explorer | System Utility Robocopy |
|-------------------------------------------|--------------|----------------------|------------------|------------------------|
| Internal hard drive, SATA                 |              | 11                   | 9.95             | 10                     |
| Deleting from hard drive, SATA            |              | 0.97                 | 0.97             | 0.97                   |
| External hard drive, USB 2.0              |              | 26.83                | 27.05            | 29.63                  |
| Deleting from hard drive, USB 2.0         |              | 1.97                 | 2                | 2                      |
| Total                                     |              | 40.77                | 39.97            | 42.6                   |

During backup of a test directory in non-archived form on a SATA drive Windows Explorer failed to copy 132 files due to the lack of support for long filenames (over 260 characters), warning the user. However, this situation means that the program has not coped with the task of copying files, though, it showed the best time of copying (Table 2).

Table 3. Time of copying (min) the test directory with a volume of 31.55 GB in archived form.

| Media type information, that receives copy | File Manager | Total Commander 8.01 | Windows Explorer | System Utility Robocopy |
|-------------------------------------------|--------------|----------------------|------------------|------------------------|
| Splitting into volumes 100 MB (301 pieces)|              | Archive creation time| 16.82            |                        |
| Internal hard drive, SATA                 |              | 6.23                 | 5.67             | 5.67                   |
| Deleting from hard drive, SATA            |              | 0.02                 | 0.02             | 0.02                   |
| External hard drive, USB 2.0              |              | 21.25                | 21.25            | 21.32                  |
| Deleting from hard drive, USB 2.0         |              | 0.05                 | 0.05             | 0.05                   |
| Total                                     |              | 44.37                | 43.81            | 43.88                  |
| Splitting into volumes 300 MB (101 pieces)|              | Archive creation time| 15.05            |                        |
| Internal hard drive, SATA                 |              | 5.6                  | 5.68             | 5.78                   |
| Deleting from hard drive, SATA            |              | 0.02                 | 0.02             | 0.02                   |
| External hard drive, USB 2.0              |              | 21.12                | 21.12            | 21.13                  |
| Deleting from hard drive, USB 2.0         |              | 0.02                 | 0.02             | 0.02                   |
| Total                                     |              | 41.81                | 41.89            | 42.0                   |
| Splitting into volumes 500 MB (61 pieces)|              | Archive creation time| 15.82            |                        |
| Internal hard drive, SATA                 |              | 5.63                 | 5.65             | 5.75                   |
| Deleting from hard drive, SATA            |              | 0.02                 | 0.02             | 0.02                   |
| External hard drive, USB 2.0              |              | 20.43                | 19.5             | 20.4                   |
| Deleting from hard drive, USB 2.0         |              | 0.02                 | 0.02             | 0.02                   |
| Total                                     |              | 40.92                | 40.01            | 41.01                  |

Total Commander 8.01 in a situation of copying without archiving showed good efficiency, but each time it met with a number of characters in the file name and asked about what to do next. This does not allow you to use it in the situation of automatic backup creation.
Robocopy [4] ideally behaved in the case of non-archived material, because without asking questions, copied everything as is and did not lose the files. In addition, this utility works in the command line mode, which gives great opportunities for its use in conjunction with the schedulers of OS tasks.

4. Discussion of results

If we compare the total time spent on the operation of creating a backup with non-archived form with the time of the backup process using archiving, we found the following patterns (Table 3 and Figure 2):

1. At first, the backup time of the non-archived directory is less than with the archive, but then the time of copying the multi-volume archive compares and even overtakes the time of copying the non-archived material.
2. The trend of decreasing the time of copying files from the volume size (segment) of a multi-volume archive clearly revealed.
3. There is a tendency to reducing the time of the archive preparation depending on the size of the archive volume.
4. Beginning with the size of the archive volume of 300 MB, there is a gradual decrease in the influence of the archive volume size on the time of copy process.
5. Robocopy showed the best results when creating a backup in combination with archiving, and demonstrated the high reliability of its work.
6. Most of all, the process of copying is influenced by the write operation on a portable disk. According to the specification (Table 3), USB 2.0 should provide a processing speed of 480 MB/s, but during the experiment, the speed changed from 280 to 380 MB/s. We assume that the reduction of the copying speed depends on the heterogeneity of the processed data such as their type, volume, etc.
7. Among the networked storages used, the most optimal in terms of price and data volume is Yandex Disk. The year of use of 1 TB of space in this storage costs approximately 3 times cheaper than in Google Drive. However, Yandex Disk does not provide the ability to control file versions, which dramatically increases the percentage of fatal errors that occur during work on projects. Therefore, in our opinion, it is better to use Google Drive as a network storage.
8. In terms of speed, network storage is directly related to the speed of the Internet channel and the amount of files. On average, with stable communication through an optical line, the amount of files equal to 31 GB should be synchronized in about 80 minutes.

5. Conclusion

Based on the results of our research, we propose a simple and economical version of the practical implementation of the backup process consisting of 4 components: the command line interpreter (CMD), the 7z archiver, the Robocopy utility, and network storage.

We also take into considering that the structure of the publisher's archive itself influences on the speed of archiving process. For example, in printing industry we can identify three parts of a common archive of the publishing house: rarely updated; constantly updated; almost not renewable.

Under the term of constantly updated, we mean the current projects of the publishing house, which ideally should be duplicated during the course of work, in the process itself, with any file change. Rarely updated areas of the publisher's archive in our opinion are software archives and a graphic data bank (clipart, etc.) which are constantly used in the work. In our opinion, they should be updated by the administrator of the network and performed no more than once in 3 months.

Project directories already containing completed orders are the almost non-renewable part of the publisher's archive. Despite the fact that as the main criterion we determined the speed of the process of creating backup copies, with a high probability of creating archive that will not be updated frequently and consisting mostly of graphic materials, compression is possible at the maximum level. The speed of creating a backup will decrease at the same time, but this inconvenience will be compensated by a decrease in the total size of the archive.
As a result, we decided that the best option would be to create backup copies, consisting of three local copies of data and two network copies. We suggest creating the first and second copies on local disks - it is SATA (internal disk), and the third copy - USB (external hard drive). Copies on Google Drive and Yandex Disk are variants of network storage. However, these directories will be on different logical disks (SATA). Local disks and network storage will contain a multi-volume archive of 7z with a splitting into 300 MB.

In this case, the Yandex Disk and external disk (USB) will be used as additional data storage media with inconsistent synchronization with the general archive.

Google Drive and the second logical disk (SATA) will act as the main tools for saving data. The SATA disk is the only media in the described system with the maximum read/write speed, and Google Disk is a tool for creating and maintaining automatic backups of files (and their versions) in the process of employees' work.

6. The source of funding
The results of the research project are published with the financial support of Tula State University within the framework of the scientific project № 2017-34». Scientific Director N.E. Proskuriakov.

References
[1] Proskuriakov N E, Yakovlev B S 2016 The improving of the effectiveness of the technology of verification of websites in the Internet 2016 Dynamics of Systems, Mechanisms and Machines, Dynamics 2016, art. no. 7819066 Conference Proceeding (Institute of Electrical and Electronics Engineers Inc.)
[2] Mosk D Work with 7zip from the command line, available at https://www.dmosk.ru/mininstruktions.php?mini=7zip-cmd
[3] Pakhomov S Comparison of 64-bit archivers WinRAR 4.2, WinZip 17.0 and 7-Zip 9.30, available at http://compress.ru/article.aspx?id=23664
[4] Malikov V We use Robocopy. Examples, available at http://vladimirmalikov.com/%D0%B8%D1%81%D0%BF%D0%BE%D0%BB%D1%8C%D0%BD%1%83%D0%B5%D0%BC-roboocopy-%D0%BF%D1%80%D0%B8%D0%BC%D0%B5%D1%80%D1%8B/