https://doi.org/10.48009/2_iis_2005_43-49

# DISASTER PLANNING FOR THE HOME USER: IDENTIFYING RISK TYPES AND PROTECTING CRITICAL DATA

**Jared Spencer, Nova Southeastern University, sjared@nova.edu**

## ABSTRACT

*The ever-growing importance of information protection and recoverability is not unique to the corporate world. Home users continue to grow more dependent on their personal systems and the data they contain. Although a large body of knowledge is available regarding disaster recovery practices, very little of it has been applied directly to the home user. This leaves the home user isolated and their data at-risk, as they typically do not have the expertise or equipment necessary to implement the techniques generally associated with corporate data protection. This research will specifically look at prioritizing data based on recoverability; and then investigate ways of providing assurance for data based on this.*

**Keywords:** Disaster Recovery, Backup, Data Reliability, Home User

## INTRODUCTION

Disaster planning is a concept that is not typically associated with the home user. However, with the continuing growth of technology in the home as well as the workplace, the importance of providing reliability and recoverability for personal data continues to grow as well. As more information is managed digitally and the home user amasses larger volumes of data that would be difficult or impossible to replace, new strategies for securing this data must be identified. This research will look at prioritizing data based on recoverability; and then investigate ways of providing assurance for data based on this.

## BACKGROUND

A great deal of work has been done in the areas of data backup and disaster recoverability. However, very little of it targets the specific issues and environment of the home user. Chervenak, Vellanki and Kurmas [1] provide a good survey of backup techniques, and also address briefly the topic of disaster recovery. In relation to the home user, however, several issues are exposed. The typical home user, for example, is not going to have access to a high-capacity tape device or other archival medium. Even assuming the presence of a recordable-DVD drive, this provides us with only 4.7 GB of recording capacity. Initiating a full backup of a 100+ GB system would prove to be a daunting task. This example may help to illustrate why "individual users almost never back up their data," as noted by Cox, Murray and Noble [2].

## METHODOLOGY

Although the research will be done with machines running some flavor of Microsoft Windows, parts of this work may be applicable to other operating systems as well. The first step will involve identifying the different risks to data, and how it will influence the approach to protection. Secondly, a structure will be created for identifying the recoverability of data. Third, the knowledge from the first step will be applied to the data of the second step to formulate a strategy for appropriate data protection. Finally, some basic tools will be implemented to test this approach. These may include existing tools used with a targeted fashion or the development of some new, simple tools that will validate this work. Ultimately, the research will show that,

given a set of identified, critical data, backup and other protection methods can be implemented to safeguard that data.

## DATA RISK IDENTIFICATION

The first step toward any type of data protection is to identify the various types of risks. Without this fundamental understanding, it is impossible to formulate an effective backup strategy. At the highest level, three types of events can be categorized: human error, system failure, and acts of God. Human errors would consist of mistakes that cause data loss. Examples would include accidentally overwriting or deleting critical data, or committing unwanted changes. The failure to take preventative action could also be considered human error, although that is a topic beyond the scope of this research. System failure can be broken down into two categories: system-wide and/or disk-level failures, and individual file errors or corruption. This distinction is important. To protect from a disk-level failure, one strategy might be to make a mirror copy on another disk. However, this approach would not protect from file corruption, as both disks might become populated with the corrupted data. Finally, acts of God would be anything that caused complete site destruction. The only measure of protection from this type of catastrophe is to keep off-site backups. The following chart (Table 1) identifies the risk types and protection measures.

**Table 1.** Risk Types and Protection Measures

| Risk Types | Examples | Protection Measures |
|---|---|---|
| Human Failure | Unwanted changes committed<br>Past data required | Multiple critical file backups |
| Disk Failure | Individual hard disk or removable disk fails | Hardware or software mirroring<br>Critical file backups |
| File Corruption | Disk sector or other errors cause file corruption<br>Unclean shutdown corrupts file | Multiple critical file backups |
| Site Destruction | Fire destroys home | Off-location storage the only protection from this catastrophe |

Critical file backups can be accomplished through a number of different techniques. Traditional file backup software can certainly accomplish the task. Other approaches may be less conventional, but offer the same level of protection. These could include copying the critical files to another computer, or saving copies on various types of removable media. The recent development of flash memory devices has shown that data can be stored in very small devices. This research will focus on testing backup techniques with a typical removable storage device.

One protection measure noted that deserves further explanation is multiple critical file backups. It may not simply be enough, as previously noted, to have a single backup of a critical file. For cases of user error or file corruption, it may be necessary to go back to a version of the file earlier than the most recent backup. This presents some unique challenges, and will be discussed further later. A backup strategy must protect the most critical, or unrecoverable data, from all of these types of failures.

## CRITICAL FILE IDENTIFICATION

The second step in this process is identifying critical files. For the sake of this research, critically important files are identified as those that cannot be replaced. Thus, the files that make up Microsoft Outlook, for example, are not critical because the software can be reinstalled from the original media. On the other hand, the .pst file that contains all of the user's email cannot be replaced, and thus is of critical importance. The loss of this file would potentially mean the loss of all of the user's email, which could represent a catastrophic event, depending on the nature of the email usage.

Identifying this distinction is important due to the limited resources of the typical home user; and perhaps one of the most significant components of this research. A common corporate backup scheme might consist of taking a full system backup once a week, followed by daily incremental backups to capture the changes from each day. However, on a modern home computer system that may typically have 100 GB or more of data and applications on it, this approach is simply not feasible. Even if the system were equipped with a DVD recordable or rewriteable drive, a full system backup would require more than 20 DVDs. Other approaches might be more reasonable, such as backing up to a second disk or creating a disk mirror. However, these approaches may not accomplish all of the requirements described earlier; and would regardless constitute more effort than the typical home user would be willing to invest. As such, a more economical and reasonable approach for the average user would be to categorize data based on recoverability; and put in place some simple processes for protecting that data. Thus, this approach does not provide any downtime protection; the user may invest significant time recovering their system from a failure. However, it does protect against the loss of unrecoverable data.

The data that is deemed to be non-recoverable can then be further broken down into categories based on change-frequency. For example, a user may have a folder that contains digital family photos. Although these could be classified as critical data, the files do not change frequently, and thus may be less susceptible to file corruption or user error. On the other hand, a file that contains financial records, such as a Microsoft Money .mny file, may change frequently. This type of file will require additional protections to ensure that several iterations of backups are available in case of corruption, etc.

For the sake of this research, it will be assumed that a critical set of data has been identified for protection. Of course, the average user may not have the level of expertise required to know how to distinguish important files from non-critical ones. Although the typical user may be able to identify the "My Documents" folder as containing important user data, certainly other files might be excluded, particularly if the user doesn't use this central location. An interesting area of research may be to find a means for automatically identifying critically important files without user intervention. But that will be beyond the scope of this research.

## PROTECTING CRITICAL FILES FROM IDENTIFIED RISKS

Now that the risks have been delineated and a framework has been outlined for identifying critical files, a structure can be developed for protecting them. As previously noted, this research will assume that a set of critical files has been identified. For this effort, the researcher has identified a test set of critical files and centralized them into the "My Documents" container. The next step will be to apply a protection framework to this container. The initial requirement for applying this framework will be to identify frequently-changing files. The researcher identified

three files from the test set (for the sake of this research) that change frequently and need to have several iterations of backups available to protect from file corruption and other errors. These are an Outlook mail file, an Outlook archive mail file, and a Microsoft Money file. All of these files are opened and used frequently, and are of critical importance. Thus, they have been identified for a further degree of protection.

Does this mean that other, non-frequently used files are not subject to corruption or possible data loss in this way? Obviously not; however, it would likely be unreasonable in the typical home environment to save multiple backup versions of every file. As such, only the most critical, frequently-changing files are being identified for multiple-copy backup. The goal is to provide at least a minimum of protection to all critical files; with the identification of those requiring multiple backup versions at the user's discretion.

## THE TEST BED IMPLEMENTATION

The test bed contained 14 top level directories and a large mixture of subdirectories and files; a total of 5,108 files containing slightly less than 7 GB of data. File sizes ranged in size from 2k to 1GB, and represented a variety of different files types, including documents, images, and presentations among many others. The diverse sizes and types of files contained in the test bed should fairly represent a critical data set of the typical home user, although certainly the number of files and volume of data will differ greatly.

In order to meet all of the protection requirements, all files must have a minimum of one backup copy, and that copy must be available off-site in the event of a site disaster. Many options are available for meeting this requirement, both in terms of hardware and software. Any solution should meet all of the requirements for critical file protection, and retain usability for the typical home user. For hardware, the researcher chose an Archos 20 GB USB hard drive. This device is portable and can be easily transported to meet the off-site storage requirement. A wide variety of other devices and media could also be used depending on the volume of critical data being protected.

There are also a wide variety of software options available. Many of these would be considered traditional backup tools, such as Microsoft's Windows XP integrated backup package. Using this software, both full and incremental backups can be taken. However, this and other similar tools are cumbersome and not always intuitive. Other simpler options can also provide effective protection. For example, simply copying the entire contents of the "My Documents" folder to the target location can provide an effective backup copy. The weakness of this approach, however, is that changes are not being tracked; meaning, the user would have to regularly copy all of the files to the target location, which can be a time-consuming process.

A more novel approach, and the one chosen for this research, is to use software that allows changes to be synchronized between two sets of files. Beall [3] recently reviewed several commercially available applications that perform data synchronization. However, for the purposes of this work, the researcher preferred a basic, freely available product. Microsoft's "Briefcase" feature, included with all recent versions of Windows, offers a free tool for this task. Briefcase allows a user to identify a set of files, and then make a synchronized replica of those files in another location. Changes are tracked, and the user can easily synchronize changes from one set of data to the other. Thus, any file changes can be synchronized over to the Briefcase

copy, allowing only those files that have changed to be updated – much like a more traditional backup package would do by using archive bits. Further, multiple Briefcases can be created allowing data to be synchronized between multiple locations, providing additional flexibility and recoverability.

## TOPOLOGY

The primary set of data (the My Documents folder) was placed initially on the Archos drive. A Briefcase copy was made on the researcher's home computer, and another Briefcase copy was made on the researcher's office computer. The My Documents folder was redirected to point to the location within the Briefcase. The process of initially creating the Briefcase copies involves dragging the files over from one location to another. Subsequently, changes can be applied by using the "Update" option. This will compare both file sets for changes and prompt the user to synchronize both data sets. An additional benefit of this approach was that it allowed the researcher to keep synchronized copies of data at multiple locations. Since the Briefcase function synchronizes in both directions, changes can be made to either (or any) set of synchronized data. By creating Briefcase copies of the identified critical files at multiple locations, the goal for off-site storage has been achieved. In fact, this goal could be achieved without the use of a second personal computer but simply by keeping the storage device offsite when replication is not taking place. The remaining requirement to be addressed is to have multiple copies of specific files.

## MULTIPLE COPIES OF FREQUENTLY-CHANGING FILES

To meet the goal of having multiple copies of frequently-changing files, several options were considered. Commercial software is available that can track file changes and backup/restore multiple file versions. However, for the sake of the test bed, a simpler approach was preferred. Ultimately, an optimal solution in this scenario would be to be able to identify a particular file or files, and have changes tracked and saved without saving an entire copy of the file. For example, several iterations of a file can be saved by simply making copies; however, for large files this can be troublesome. If the user has a 250 MB database file, and chooses to retain three backup revisions of the file, this brings the space required for this single file up to 1 GB. However, for the simplicity of the test bed, this type of approach will be used.

Multiple backup versions of the three identified files were created. The file names were updated with extensions save1, save2 and save3. Although this approach is rather simplistic, it will be effective in validating the research. At periodic points, the files will be updated by copying the original file to the save1 extension, the file with the save1 extension to the save2 extension, and so on. After several iterations of this process, four unique files are created. Each of these files is later replicated using the Briefcase. This additional step has now created multiple copies of the identified frequently-changing files. However, it would be quite cumbersome to manually copy the files from one version to another. To automate this process, the researcher created a simple batch file to perform the copies automatically. To minimize the overhead of this task, the batch file checks to see if the file has changed. If it has not, the process of updating all the copies is skipped. If the file has changed, all copies are updated. The batch file process was placed in the shortcuts for starting the applications to cause it to run each time the respective application was opened. Another way of automating the process would be to create a scheduled task to have it run at given intervals, which might be a good solution depending on the type of file being

copied. For this research, the files being used are locked when open by the application; as such, it was necessary to perform the copy when the application was not running.

The test bed should now meet all of the goals for critical file protection. A methodology has been established for having at least one backup copy of each file, and that copy can be held offsite in the event of natural disaster. Further, a structure was developed to create several copies of the most critical, frequently-changing files to protect from file corruption and other errors. The next step will be to validate the test bed by blowing up some data!

## RESULTS

The first set of testing was to introduce human error. For this test, the researcher first chose several files and deleted them. Next, a Word document was updated with some unwanted changes. Finally, some data was removed from an Excel spreadsheet and then saved. All of these actions are typical examples of human error. The test bed recovered flawlessly from each of these problems. Although the use of Briefcase worked well for the sake of our research, it may not be an optimal solution for end-customer use; unless the user was skilled and knowledgeable in file management. For example, after the files were deleted, upon initiating the next synchronization, Briefcase assumes that the deletion was intentional, and will attempt to remove the synchronized copies. The user will have to identify that in fact the opposite needs to occur and the deleted files need to be recreated. Although this is a simple Briefcase task, it does place the burden on the user to properly manage the situation. However, it would also be worth noting that the same would be true of a more traditional backup; if the user only had one backup and overwrote it with the files having been removed, they could no longer be recovered.

The second scenario noted earlier is a disk failure. For the sake of this research, we will consider this failure in the same context as a site disaster, which will be covered later. The third scenario involves file corruption. In this case it will be assumed that a hardware error has rendered the test file unusable. This can occur from disk problems, unexpected power failures and other related problems. To test this scenario, one of the frequently-changing files in the test bed (an Outlook .pst file) was modified in a text editor and some characters were randomly added to the file. This has the net effect of corrupting a non-text file. Upon attempting to start Outlook, the program refused to open the identified .pst file. At this point, the steps for recovery depend upon when the corruption occurred. If the file became corrupted after the latest Briefcase synchronization, another update can be initiated to replace the corrupted file with the good copy. If the corruption occurred unknowingly prior to the last replication, then one of the .save files must be restored manually.

Scenarios one and three were both tested successfully. In the case of not knowing when the corruption occurred, recovery may involve a series of steps to determine the best good copy available. No automated tools were written for the sake of this research to restore the .save files. In the case where all Briefcase copies were corrupted, the .pst file was manually copied from a .save file. In this case, any email activity since the last save was lost. This scenario demonstrates the importance of having multiple copies of frequently changing files; and in some cases, it may be desirable to have multiple copies of files that don’t necessarily change frequently, but are of the highest importance. File corruption can impact any file – not just those that change frequently.

Arguably the facet of critical data loss that fewest home users are prepared for is a site disaster. Although the research shows that few home users backup their data, likely far fewer keep off-site backups. This research shows, however, that keeping an offsite backup does not have to be a difficult proposition; and in fact this may be one of the more valuable discoveries of this work. To test site disaster recovery, the researcher turned off his primary home computer for one week, and worked on a secondary home machine starting with a blank disk. In this case, it certainly would be beneficial to have a full system backup for easier restoration. However, the premise of the research is that non-critical data can be restored. So the test machine was built from scratch with Windows XP and loaded with the software required for the test bed. No files were loaded that were not part of a standard media distribution or freely available for download.

The results of the testing were successful. Although a time-consuming process, the researcher was able to reload the necessary software and restore all critical data by creating a new Briefcase on the test computer. Some configuration issues were encountered, such as changing the application settings to store and search for files in the appropriate location; all of which demonstrated that this type of protection will require significant effort on the user's part to recover from a true disaster. However, no critical data was lost, which validates this method of recovery.

## CONCLUSION

The testing results clearly validate the research theory that home users can effectively and economically protect their critical data. By employing some basic tools and strategies, the researcher was able to successfully protect a set of test data from all identified causes of data loss. Like previous research efforts, such as those by Cox, Murray and Noble [2], this work demonstrates that non-traditional means can provide effective data protection. The primary factor that distinguishes this research is in the notion of identifying and protecting only critical data rather than the entire computer system. By taking this novel approach, the scope has been sufficiently reduced to make it more feasible for the home user to protect their critical data.

Several potential areas for continuing research can be, and have been noted. As a means of validating the research, Briefcase proved to be an acceptable tool. However, a solution more specifically tailored to the specific goal of file protection would likely prove more effective. To extend that idea even further, a complete package solution that offered this functionality along with a cleaner implementation of multiple-file backups would be an exciting avenue for further research. Finally, developing an assisted process to guide users through the process of identifying critical files, or automating it entirely, would be an interesting research goal. The researcher envisions a solution that encompasses all of these requirements: automating the process of identifying critical files, providing a robust means for their backup, and incorporating the additional functionality of multiple backup protections for individual files.

## REFERENCES

1. Chervenak, A., Vellanki, V. & Kurmas, Z. (1998). Protecting file systems: A survey of backup techniques. *Joint NASA and IEEE Mass Storage Conference*.
2. Cox, L.P., Murray, C.D. & Noble, B.D. (2002). Pastiche: making backup cheap and easy. *ACM SIGOPS Operating Systems Review*, *36*(SI), 285-298.
3. Beall, R. (2004). Keeping Your Data In Sync. *Network Computing*, *15*(4), 93 - 96.