

Backup Strategies

When people ask me about backing up their data, or when people tell me that they've lost data because of an event, I'm reminded of an anecdote that Nicholas Negroponte tells in his *Being Digital*, probably the most influential book in my personal history, the book that actually got me interested in computers and high technology. He relates a story of visiting a high-tech firm's corporate offices and is asked to declare the value of his old Apple PowerBook. He estimates the value as between one and two million dollars; the receptionist, by contrast, looks it over and declares the value as about \$2000 - approximately the price Negroponte paid for it when he purchased it. The anecdote illustrates the problem of "bits": Negroponte valued the information contained within far more than he valued the actual atoms that made the reading of the data within.

And herein lies the basic problem - how does one create and value a backup system? How much is your data worth? Certainly, individually, no piece may be worth that much - that one cute picture of your pet, or that one document you submitted may not be worth the money of a whole (potentially complicated backup system), but what you're externalizing in this process is your time and the value of memory. Once you begin to internalize such costs - Can I get a picture of my kid at age 3 again? Can I afford to spend another two days recreating this document? - then the apparent sticker shock of a backup system begins to seem less like a cost and more a form of insurance.

A backup can be a complicated beast, when in fact it should be one of the easiest things in the world to do. What kind of backup you choose depends on a number of factors:

1. The purpose of the backup.
2. Your tolerance for data loss.
3. The amount of time you want to dedicate to the process.
4. The total amount of data you have.

Depending on what your answers are, the backup can be as simple as sticking in a CD and making a copy of your files once every few weeks, to several hundred dollars of open source, multiply redundant off site storage. I'll review each in turn, because I feel this is the order these questions should be asked. Other people have different criteria; for me, the purpose and tolerance is most important, and everything else follows.

The Purpose of a Backup

Backups exist for many reasons, driven by the same underlying goal: to reduce data loss. Why data loss occurs determines why backups need to be made.

Case 1: Catastrophic hardware failure. Ask the average person on the street why keeping regular backups is beneficial will likely elicit the "I'm afraid of a hard disk failure" response. Certainly, hard drives - and other computer components - fail with alarming regularity. To prevent against the loss of data, backups must be designed to withstand component failures not only in the primary copy, but also the backup. Backing up your data to a copy that turns out to be bad when you need it is a worst case scenario, so consider multiple backups. Moreover, rarely is catastrophic hardware failure or theft or loss predictable; therefore data must be in a format that is readily readable - and on a completely different device altogether. The best way to protect against this sort of failure is to maintain a copy of your data on more than one device and, ideally, on more than one form of media.

Case 2: User error. In my experience, most data loss is not caused by hardware failure or loss: it's caused by a moment of carelessness. Whether it's deleting files you had thought you copied, or entered your changes into a file and forgot to save them, you are far more likely to lose data because of a moment of carelessness than because of a hardware problem. Protecting against this sort of error is both time

consuming and, more often than not, is resource intensive and requires support from the operating system. That's not to say it can't be done, but it's harder to fix stupidity than a hard drive.

With the newer major operating systems – Microsoft Windows Vista and Mac OS X 10.5 “Leopard” – this functionality is built in as “Shadow Copy” and “Time Machine”, respectively. While such a system is not built in to Linux, nearly equivalent functionality can be provided through the “Time Vault” package available from a repository near you.

Case 3: Future-proofing. It is a fact of life that technology that is standard and widely used today will be gone tomorrow. To prevent against data loss because of obsolescence, there are many, many strategies, but all of which require careful planning and research.

Broadly speaking – keep very important data in multiple formats, including paper if need be; keep a machine that is able to read your data around either in physical form or virtual form; and, most crucially of all, avoid digital restrictions management (DRM) technology. A fundamental step that will help here is to ensure that your backups are not in a proprietary form; as I learnt the hard way, despite the name and interface being the same between the Windows 98 and Windows XP backup utilities, they used different, incompatible formats that nearly rendered the backup unusable.

Second, it's worth keeping all your files in a form that is universally readable and preferably in an open, standards-based format.

A third approach is to keep virtual machines available which are capable of opening the original operating system and application combination that you need. So for example, if you need to access a document that was created by an application like WordStar, it may be worth your while to create a virtual machine with DOS installed that you can run if need be.

Keeping your files up-to-date with the latest formats is another way of protecting against obsolescence – usually newer versions of software can read and write to at least a couple of prior iterations, but it's unlikely that a 2008 release of a particular software will be able to write to a 1993 version, while the 1997 release can read the 1993 version, and the 2008 version can read the 1997 version.

Finally, keep in mind the fate of the 5.25-inch floppy and remember to update the hardware interfaces too as you go along. A good bet would be to evaluate the interfaces about every ten years. Ten years ago, parallel ports were dominant and today USB is dominant – but a few machine still have both and can provide a bridge between the interfaces as a last chance to move such data.

Case 4: Loss of property. Whether the cat dropped the iron and caused a fire, or whether your phone was stolen, it is always important to have copies of the data in multiple places. Quite often, this is called offsite backup, and if there is very important data, it may be worth keeping some data offsite. For very important documents, you may want to consider keeping another copy with a family or friend who is located in a different place from you. It's simple enough to make a single copy and take it over when you go. If you trust online services, it may be worth investing in such a service as well.

Tolerance for Data Loss

No matter how well designed your backup system is, it is inevitable that some data loss will occur. The chief goal of a backup system is not to prevent the loss of data, but to minimize the impact of the data loss. Broadly speaking, data breaks down into the following categories:

Irrelevant – data that is either junk or freely and easily retrievable from other public sources. This includes things like most temporary data files, browser cache, most applications and application installation files, and such.

Relevant – data that is available from public sources, sometimes for a fee, and can be reacquired through those public sources. This includes things like music you may have downloaded, or electronic books, or receipts and statements. If you are using IMAP, then most email would indeed fall into this category, as would telephone numbers.

Important – data that is created by you and is not generally available from public sources, but the loss of which will not end your world as you know it or will not take much time to redo or recover. This would include photos, pictures, home videos, and documents that are not line of business, and email on servers that cannot be re-downloaded, such as POP servers.

Crucial – data that is created by you, is not available from public sources and is work that cannot be redone. This category usually only includes current work or current projects that are time-bound or cannot be redone.

The vast majority of data, I find, falls into relevant and important. If you're keeping a good clean computer to begin with, there will rarely be much irrelevant data; similarly, very few people are doing many individual projects so very little data will be marked as crucial. Generally speaking, I try to backup relevant data once a month, important data once a week and crucial data every day, if not more often. Irrelevant data usually gets backed up once – when I am setting up a computer to begin with, I often take a snap shot of how it looks in case I need to quickly bring a computer back to using those applications I have installed and with those settings that I have invested time in.

Time Constraints

As in many other activities, there is a tradeoff in terms of time spent and backup ability. If you're willing to spend lots of time setting up and using your backup system, you're less likely to lose data. If you're less willing to spend time on setting it up, then chances are you're not going to have reliable and extensive backups. Depending on your risk aversion, you will naturally compute out what the optimum balance is for yourself. The second time tradeoff with respect to backups is in initial setup time versus time spent on each individual back up session. If you're willing to spend time to set it up perfectly, you should be able to forget that the backup system exists until you need to use it to retrieve a file from it. On the other hand, it's easy enough to just drag and copy all your files, but it takes more time to do so every time you make a backup.

My current system is setup to backup everything according to a schedule to a set of external hard drives. This is kept in sync with a second drive of exactly the same size, so effectively I have two copies of the backup – what's called a RAID 1 setup in technical parlance. I can lose one of the hard drives and still have a complete intact backup. I also sync all my USB flash drives through a script that watches for the sticks to be inserted and then copies the contents to a folder. While I could subscribe to a number of online backup services, I decided a long time ago that my data is safer and more secure in my possession. Thus I just encrypt and backup the absolutely crucial things I want to keep onto a DVD every so often and leave it at a family member's place, or at my work place. Make sure to write "Do Not Throw Away – Contact [Your Name] First" to ensure that it avoids the fate of unmarked discs everywhere. You may benefit from putting a reminder on your calendar to do so every so often.

Data

How much data you have will determine the optimum backup method. For the vast majority of people, all the important data they have will fit comfortably onto a DVD. Others may find hard drives more

appropriate. The table below lists the maximum capacity one can get from each type of medium and what the approximate price ranges are for each type of medium.

Medium	Capacity	Cost	Notes
CD	650 – 700MB	\$0.05/disc + \$25 writer	Given that most computers come with a CD burner, this is a negligible expense and quite convenient. For those with limited number of documents, this is the easiest and cheapest way to go. You also need to take care to ensure the disc is not damaged through scratching. Some discs are also known to corrode over time, so pick reliable media, test them every so often and store carefully.
DVD	4.7 – 9.0GB	\$0.10/disc + \$40 writer	As with CDs, many computers come with a DVD burner. Essentially the same caveats as CDs apply, but the space on each disc is about an order of magnitude larger.
BD-ROM	25GB	\$10/disc to \$25/disc + \$450 writer	Same caveats as CDs and DVDs, and about five times the space of DVDs. However, the burners usually need to be bought separately, and Blu-ray is in a fight with the rival HD-DVD format, which may mean that your investment will be for naught, and your media unreadable. However, media is generally of the highest quality, though painfully slow.
USB Flash Drive	128MB – 32GB	\$5.00/drive to \$400/drive	Depending on the amount of data that needs to be backed up, this may be the easiest way to make a complete duplicate. While the process is mature enough and the suppliers of the chips are few enough that the quality of a name-brand and a non-brand is basically the same, make sure the connector is solidly attached.
External Hard Drive	20GB – 1TB	\$40/drive to \$400/drive	External hard drives suffer from the same problems as internal hard drives – they are relatively fragile and careful handling is a must. However, they also come in excellent storage for price metrics and it may be cheaper to buy a bunch of hard drives and keep identical copies. They are also generally less portable than are flash drives.
Storage Robot	80GB – 4TB	\$450/unit + \$40/drive - \$400/drive	Drobo is the first of a line of companies that have come out with so called “storage robots” that are basically dedicated devices that simplify and automate backup. While the price is high, the convenience is unbeatable and the device is easily expandable. The only question is that of cost – assembling a dedicated server costs anywhere between half and a quarter of buying and outfitting a Drobo with equivalent storage capacity. But your time outlay is correspondingly less.
Storage Server	Varies.	Varies.	A home storage server can be as simple as an old, Linux-based computer that has a bunch of hard drives attached that it shares out, to a top-of-the-line “prosumer” or small business setup with multiple redundancies and power backup. With increasing cost comes increasing reliability and depending on your budget, it may be possible to assemble a backup system that rivals that of Fortune 500 companies. But do you really need that kind of backup system?

Cloud Storage	Unlimited	Free - \$0.20/GB every month	So called “cloud storage” is storage that is provided by an offsite provider, like Amazon’s S3, or X-drive or even through Gmail with the help of appropriate extensions. The advantages of this is that essentially you do not have to worry about providing every gigabyte yourself, nor do you have to worry about redundancy – your provider will (hopefully) take care of such matters. And as long as you have access to an internet connection, you can access your data. However, what you gain in ease of use and convenience, you give up in privacy and control of your data. While you can hope providers are reliable and will take steps to keep your data safe, it takes a single breach to compromise your data. There is also the question of keeping your data under your control – while heavy encryption can generally ensure that your data is all but irretrievable without your say-so, the potential exists to have your data accessed and worked upon without your knowing it. There is also the question of what happens to your data is for any reason your monthly payment is not processed or received. Finally, there is the question of speed – depending on how fast or slow your internet connection is, it may take several minutes to several days to upload and download larger amounts of data.
---------------	-----------	------------------------------	--

Summing Up

What are the take away lessons from this long article? Well, if you are very paranoid about your data, you’d pick multiple backup systems, backup everything often and spend lots of time and money on your backup system. That unfortunately means that you’re also likely not to use it, because it’s very intrusive and time-consuming.

In my opinion, the best system has the following properties:

- Requires very little conscious thought once it is setup.
- Survives at least two failures – it should allow for the failure of one computer and one backup.
- Allows you to be able to access your data at least ten years after you backup the data.
- Allows you to recover from a stupid mistake at least a week after you make it.
- Is tested routinely – at least once a week.
- Keeps absolutely the most critical data offsite in addition to an onsite backup.