Preventing Data Loss Through Prudent Archiving

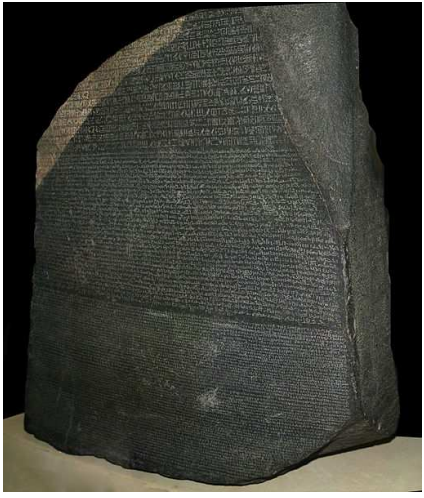
BRUCE MOMJIAN

Prudent archiving requires explicit risk analysis to limit exposure to possible data loss. Some common sense archiving rules can greatly reduce the chance of getting that sinking feeling when data is irrecoverably lost.

https://momjian.us/presentations

Last updated: August 2022

An Ancient Archiving Example: The Rosetta Stone

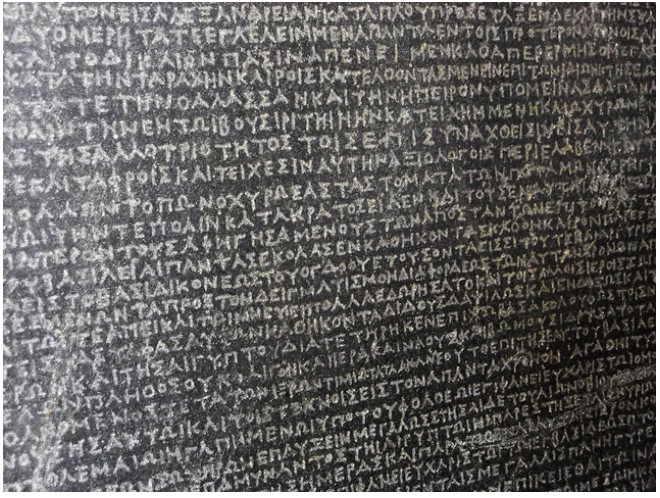


https://commons.wikimedia.org/wiki/File:Rosetta_Stone.JPG

Rosetta Stone History

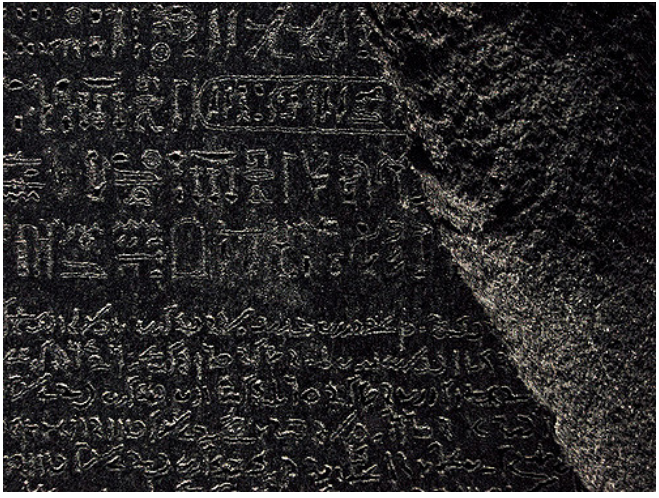
- Created in 196 BC in Egypt by Ptolemy V
- Records tax reductions and temple statuary rules
- Discovered by the French in 1799
- Decree is recorded in three languages
- Eventually enabled the translation of Hieroglyphics

Rosetta Durability: Oops



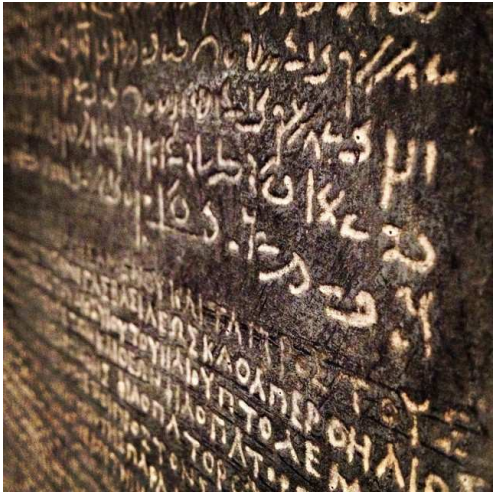
https://www.flickr.com/photos/zongo/

Rosetta Allows Reading of Hieroglyphics



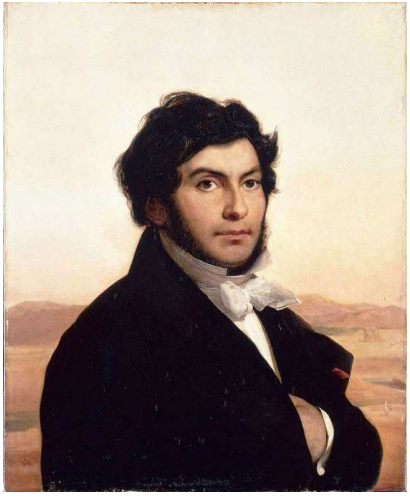
https://www.flickr.com/photos/bradworks/

Three Sections, Three Languages



https://www.flickr.com/photos/zouny/

Hieroglyphics Translated 24 Years After Its Discovery



Jean-François Champollion, by Léon Cogniet

Archiving Lessons From The Rosetta Stone

- Storage durability
- Information usability, i.e., just because you have the file doesn't mean you can understand it

Storage Durability

- Media lifespan
- Multiple copies
- Distributed copies

Media Lifespan

- Stone, clay
- Paper
- Paper tape, punch-cards
- Floppies
- Hard disk
- CD/DVD
- Tape
- Memory stick / flash

Be prepared to migrate all data to a new storage media format every few years.

My Backup Media History

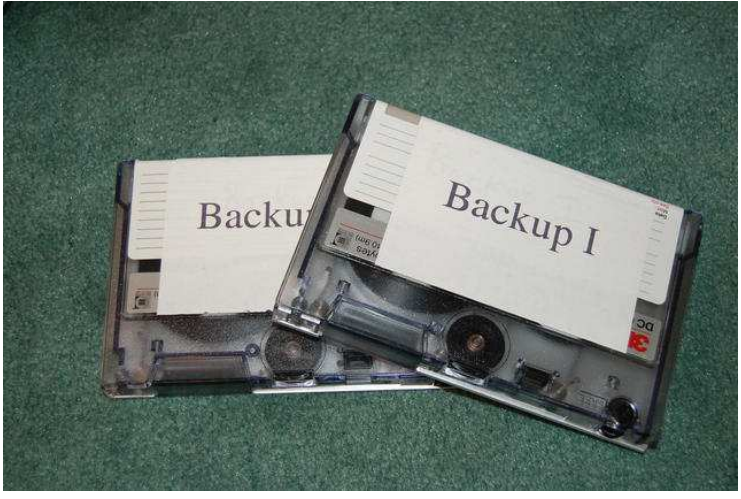
https://www.youtube.com/watch?v=AvXXkB2jicO

https://www.youtube.com/watch?v=e7v13Zu3SXc

DEC Disk Pack: Early 1980's, 10MB

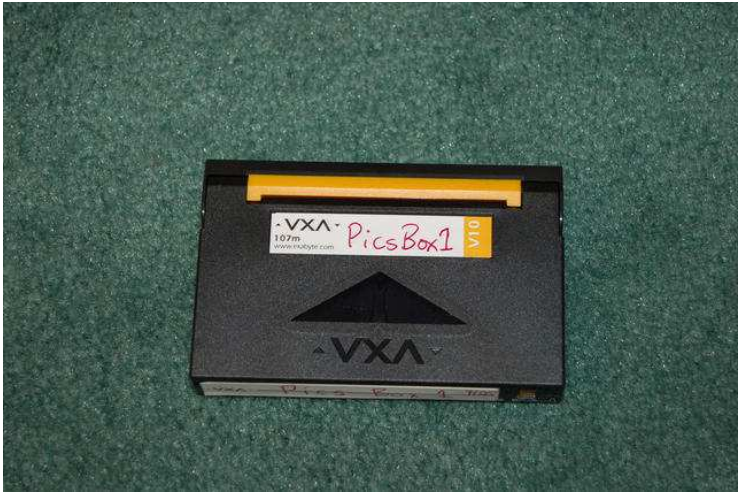
Floppies: Late 1980's, 360KB

QIC 150 1/4” Tape: Early 1990’s, 150MB

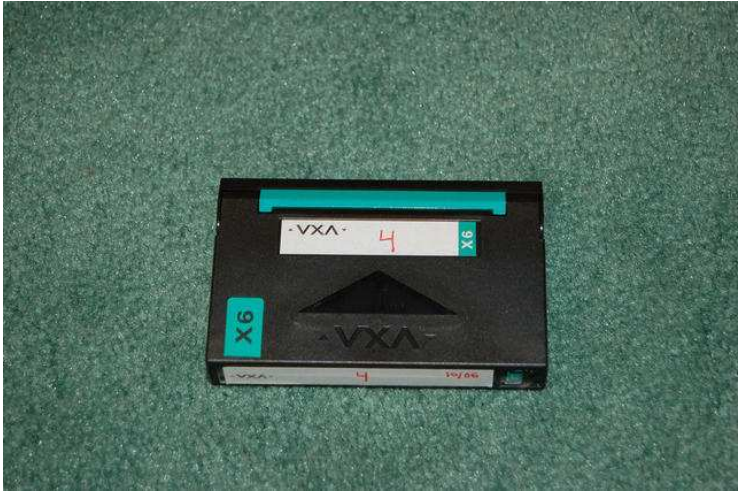


DAT Tape: Mid-1990's, 4GB

VXA-1 V10: Early 2000's, 20GB



VXA-320 X6: Mid-2000's, 40GB



VXA-320 X10: Late 2000's, 86GB

Seagate Barracuda SATA Disk Drives: 2008, 1TB

Three SATA Disk Drives

Multiple Distinct Copies

Multiple distinct copies offer protection from:

- Corrupt files
- Accidental deletion
- Multi-device failure

RAID only protects from single-device failure.

Distributed Copies

Distributed copies protect from:

- Theft
- Fire
- Flood
- Massive device failure (electrical surge, lightning)

Protect your backups; they contain all your data. They should not be lost or read by others.

Data Safe

Inside the Data Safe

Make sure the temperature tolerances of the safe are acceptable for the media stored in the safe.

Safe Deposit Box

https://www.flickr.com/photos/ell-r-brown/

Verify Archives

- Regularly restore a random file and check its validity
- Regularly list the contents of an archive and verify that all file names appear

Ideally verification is both automated and manual.

Is the Device Based on Open Specifications?

- Physical connection: 80-pin SCSI, mini-USB, SATA, DVD+R
- Command set: SCSI, USB
- File aggregation: ext2, FAT32, tar, cpio

Sometimes multiple formats are involved, e.g., a CDROM written using ISO 9660 (CDFS), inserted into an CDROM drive that connects to the host computer using an IDE interface. Closed device specifications can make old archives more difficult to restore.

The Archive Usability Problem: Can You Use the File?

Can you read a Wordstar file today? Many files are useless without the applications that created them, but often the software, operating systems, and hardware needed to run the applications are no longer unavailable.

Usable Formats

Consider if all archived files are usable:

- documents
- images
- audio
- videos
- databases

Audio provides a good example that changing formats require regular updates: reel-to-reel tape, cassette, audio CD, MP3 memory stick.

File Format Upgrades

Just as it is necessary to upgrade to new storage media format every few years, so it is necessary to upgrade to new file formats every few years so that existing applications can use the archived files. This wasn't necessary for stone, paper, etc., because both the media and format had long-term durability and usability, e.g., a 200-year-old book, but computer files to do not offer the same retention lifespan.

Open Standards

Storing files using open standard formats increases the probability of future file usability; archiving the source code used to create the file also helps; see https://suchanek.name/texts/archiving/ for more details.

Historical Archive Retention

Historical archive retention allows recovery from:

- Failed backups
- Corrupt files
- Unintended changes

What is the cost of retaining old backups vs. the cost of data loss that is detected only after all historical backups are gone?

Incremental Archiving

- Incremental archiving can reduce storage costs
- Deletion and renaming of files can be problematic
- Managing and restoring incremental archives can be complicated and error-prone

Multimedia Files

- Often created in batches, e.g., dumping data from a digital camera or video camcorder
- The size of multi-media files often requires different archiving methods

Archive Suggestions

- Consider preserving file modification times as well as file contents
- Store the list of archived files separately for easier file restoration
- Creating a well-defined area that is *not* archived often reduces archive requirements without increasing the chance of data loss

How Good Has Retention Been In The Past?

What is your oldest saved file? Past retention is a good indicator of future retention success.

My Oldest File

```
%TITLE "SCREEN SWAPPING TSR"
;this program initializes a 4k buffer to allow the user to switch between
;several screens while using the computer. "

PRINT_SCREEN_NUM   = 05                    ;interupt 05
TOP_OF_RAM_SEGMENT     = 0a000h
MONO_SCREEN_NUMBER     = 7h

SIZE_OF_SCREEN_BUFFERS          = 80*25*2 ;80 columns, 25 lines, 2 attributes

include "BIOS.INC"
include "DOS.INC"
include "KBD.INC"                       ;in \assemble\include
...
xchg_next:
      sub bx,2                            ;last word
      mov ax,[es:bx]                      ;get current char
      xchg ax,[ds:bx]                     ;exchange with extra
      mov [es:bx],ax                      ;put extra into current
      cmp bx,0
      jne xchg_next                       ;at bottom of buffer
```

My Archive Regemin

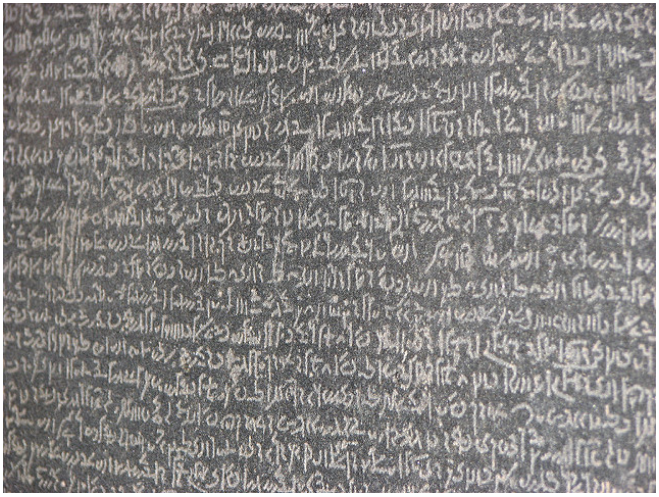
- Backup every night to separate file system and tape
- Overwrite tapes every night, cycle tape every week in eight-week rotation
- Create backup of *changed* files every night; retain for one week
- Deleted files are either on the previous tape or archived as *changed*
- Video is archived to three SATA drives, periodically synced up and stored in separate locations

Lessons

- Be prepared to upgrade to newer storage formats
- Be prepared to upgrade to newer file formats
- Keep multiple, distributed archive copies
- Retain some historical archives
- Prevent archive loss and theft
- Verify archives regularly
- Incremental or special multimedia archiving rules should be considered carefully

Primary data loss will happen; it is just a question of when. Archiving allows recovery when loss happens.

Conclusion: Ramses Name In Hieroglyphics



https://www.flickr.com/photos/bortescristian/