Information

- This is an in-class exercise developed by Digital Preservation Southampton at the University of Southampton’s Digital Humanities team.
- In honor of World Digital Preservation Day, we are sharing a public version of this exercise for educational use by educators and trainers.
- All resources mentioned are available on the Zenodo repository. All files used come from the DiscMaster website.
- If you have any questions, please contact digitalpreservation@soton.ac.uk.

Integrity Checking Detective

In-Class Exercise

World Digital Preservation Day 2024

Digital Preservation Southampton

Integrity Checking

- In the preservation of digital cultural heritage, integrity checking is crucial to ensure long-term access.
- Integrity checking is the process by which digital preservationists routinely check that the information inside a file has not been corrupted or modified.
- Files can become corrupted through several processes including:
  - Bit Rot
  - Malware or cyberattacks
  - Transfer errors
  - File format obsolescence
- When implemented along with regular backups, integrity checking can mitigate risks of data loss.

Checksums

- One main strategy to implement integrity checking is the use of checksums or cryptographic hashes.
- Checksums are unique strings of alphanumeric characters that represent the contents of a file.
- If a file changes or becomes corrupted, its checksum will completely change.
- There are different algorithms to create checksums, e.g. SHA-256, MD5.
- Checksums look like this
  - MD5: 6f40762cfabb8d780eb9fdff3ea66b89
  - SHA-256: BD94760347BABBB0B12ADFEB41FF01B90DD7F4C16F9B6C2088CD2526F6223898

Verifiable File Manifests

- When ingesting a digital collection, preservationists should (ideally) receive a verifiable file manifest which includes file-level metadata, such as checksums.
- This inventory constitutes a reference point for preservationists to work with during processes such as integrity checking.
- There are tools that can create automated verifiable file manifests, such as DROID.

Backups

- You may be wondering, what if a file is corrupted?
- If a file is found to be corrupted, it can be restored from an existing backup.
- Depending on the situation, the file may also need to be migrated to a more suitable file format or may require an emulated environment for access. These preservation methods exceed the scope of this exercise.
- Ideally, there will be existing back-ups of the heritage files including:
  - Copies on a local computer.
  - Copies on a hard-drive or disc, stored in a location with a different environmental profile to safeguard from environmental risk.
  - Copies on the cloud.


YOUR TASK

(In groups of 3 or 4 with 1-2 screens)

You are a team of preservationists at a cultural heritage organization, and you have just received a file transfer from a depositor. This is a set of a few different files of different formats, some of which are rare. You suspect that some of the files may have become corrupt during the transfer. Thankfully, the depositor has also provided cloud backups of the files, so you will be able to restore them.

But first, you need to **identify the corrupted file(s).**

What you received from your depositor

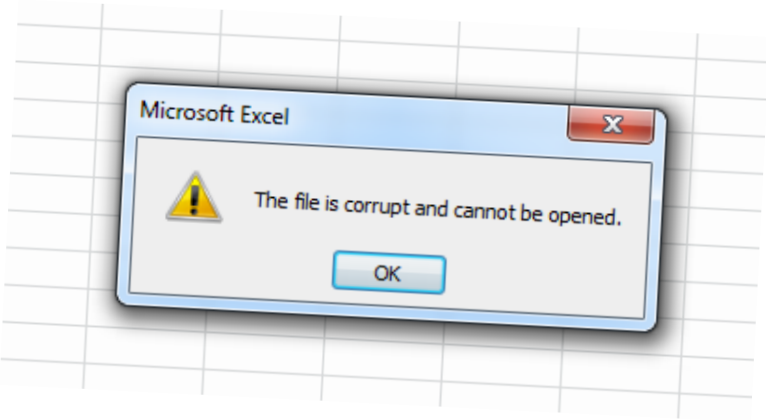
- A folder with files to be ingested (exercise_data_corrupted.zip).
- A verifiable file manifest created with DROID including the original checksums (exercise_data_vfm.csv)

Download and unzip these files and place them in a sensible location on your computer.

ID	PARENT_I	URI	FILE_PATH	NAME	METHOD	STATUS	SIZE	TYPE	EXT	LAST_MOI	EXTENSIO	SHA256_HASH	FORMAT_	PUID	MIME_TYP	FORMAT_	FORMAT_VERSION
2		HIDDEN F	HIDDEN F	exercise_data		Done		Folder		2024-08-2	FALSE						
3	2	HIDDEN F	HIDDEN F	00052_Bit	Signature	Done	3605	File	png	2024-07-3	FALSE	10244fc64fce6b212d778c753ab4cbc87f247eabdcf82094f3fc41560e9f886a	1	fmt/11	image/pn	Portable N	1
5	2	HIDDEN F	HIDDEN F	BUSH.FLI	Signature	Done	656542	File	fli	2024-08-2	FALSE	16dee398cc5b4d9cce97a977f90cc9e81b346e1b91d2f9f0feb1850beecfcf92	1	x-fmt/154		AutoDesk FLIC Animation	
9	2	HIDDEN F	HIDDEN F	Black_Dog	Signature	Done	7317021	File	mp3	2024-07-3	TRUE	7ac0a57e04fbbb84ec2816befa4cc2645760496b9f5247a1d277270d19ffeb08	1	fmt/1812	audio/aac	Audio Data Transport Stream	
8	2	HIDDEN F	HIDDEN F	X-FILES.AU		Done	360470	File	au	2024-08-2	FALSE	b98e456fdcc19a66aa85b7795e0587fa2c08e985c6c6e07caeda21c701891c58	0				
4	2	HIDDEN F	HIDDEN F	alt.fan.bo	Signature	Done	8091	File	bonzo-do	2024-07-3	TRUE	499196723fe67e17a8f389d4223ffd6697430489dd33691700b3514dae23b365	1	fmt/950	message/	MIME Ema	1
6	2	HIDDEN F	HIDDEN F	clinton.gif	Signature	Done	19955	File	gif	2024-08-2	FALSE	44240c709051c02b8d381361066cc0e206689d4d8e171b09a3b81aa6b3a84454	1	fmt/3	image/gif	Graphics I	87a
7	2	HIDDEN F	HIDDEN F	shellnew.	Signature	Done	11376	File	lwp	2024-08-2	FALSE	20d19c341c0fd3513cb372d5451417ea090b5a746cd93073cbd29d56da86300c	1	fmt/340	applicatio	Lotus Wor	97/Millennium

KEEP IN MIND

- You may have come across corrupted files before. Usually, this happens when you try to open a file, and your program of choice informs you that it is unable to open it. Therefore, you may be tempted to try to open all files in the folder and see which work.
- However, this workflow is not sustainable nor scalable. You may simply have too many files to check for. In this dataset, some files have rare formats, and your computer may not have the appropriate programs to open them.
- **The fact that you can't open a file does not mean it is corrupt.**
- Furthermore, sometimes you can modify a file's contents **just by opening them.**
- Therefore, it is better to work with batch processing tools, such as checksums.



WHAT TO DO

- There are various online tools that can create a checksum for a file.
- This resource can create SHA-256 checksums https://tinyurl.com/sha256checksum
- Upload the files directly from your computer into the tool and compare the generated checksums with the ones on the verifiable file manifest.
- Identify the corrupted files and make note of them.

Having trouble? Here's a hint

- Your colleague, who is trained in DROID, was able to create a file register of the files you received (exercise_data_corrupted_vfm.csv). You are now able to compare this to the original file manifest. Can you spot the differences?

Once you have identified the corrupted file(s), you can restore them from a cloud backup of the files (exercise_data_backup.zip). Download the relevant files and replace them in your computer folder. Your collection is now restored!

Then, make note of this transfer in the Verifiable File Manifest (exercise_data_vfm.csv). Add a new column called 'Fixity' to record your actions. Some of the details you may want to make note of include:

- Date/time of the recovery.
- Details of the corruption, such as how it was detected.
- Agent/entity that executed the recovery (person or software).

Reflection Questions

In your groups discuss:

- Why does file integrity checking matter for people who care about heritage?
- What is the importance of working towards the long-term preservation of digital cultural heritage?
- In what ways could file corruption occur in a digital preservation environment, and how can we mitigate these risks?

Further Reading

In a cultural heritage setting that conforms to PREMIS (Preservation Metadata Implementation Strategies) Standards, file integrity check failure and file recovery from cloud backup would be recorded as two separate **events** associated with the fixity of the object. These events would be recorded in adherence to the **semantic units** described by the PREMIS standard, ideally in XML format, that would then be attached or associated with the object record.

You may learn more about the PREMIS standard here https://ufs.libguides.com/c.php?g=1113411&p=8118665