Dataset Plan for NTFS vs EXT4 Performance and Forensic Recovery (USB/External Drive Environment)

This document outlines the datasets that will be used for the project **A Dual Perspective on NTFS vs EXT4: Performance and Forensic Recovery**. Since the primary environment involves a USB/external drive, datasets have been selected to represent typical portable storage workloads including small text files, images, videos, and mixed document types. Both open-source datasets and synthetic files will be used to ensure reproducibility and relevance.

1) Small Files (Documents, Logs, Configs)

- Enron Email Dataset (small text files for metadata + recovery) - https://www.cs.cmu.edu/~enron/
- System Log Dataset (LogHub for realistic log files) - https://github.com/logpai/loghub

2) Medium Files (PDFs, Images, Mixed Content)

- Project Gutenberg (public domain books in TXT/PDF format) - https://www.gutenberg.org/
- COCO Dataset (image files for recovery testing) - https://cocodataset.org/#download

3) Large Files (Videos, ISOs, Archives)

- Sample Videos (MP4, MOV, AVI test files) - https://sample-videos.com/
- Linux ISO Images (large OS files for throughput tests) - https://ubuntu.com/download

4) Mixed Portable-Storage Workloads

- GovDocs1 Dataset (documents: DOC, XLS, PPT, PDF for forensic testing) - https://digitalcorpora.org/corpora/files
- Digital Corpora Disk Images & File Sets (realistic user data for forensic recovery) - https://digitalcorpora.org/corpora/disk-images

5) Synthetic Data Generation (Reproducibility)

In addition to public datasets, synthetic test files will be generated to simulate controlled workloads. These files will be created directly on USB/external drives to measure read/write speeds and recovery capabilities:

- Linux: dd if=/dev/urandom of=usb_test_large.img bs=1M count=1024 # 1GB file
- Windows PowerShell: fsutil file createnew usb_test_large.img 1073741824 # 1GB file

These datasets collectively provide a realistic and reproducible workload for evaluating file system performance (NTFS vs EXT4) and forensic recovery in USB/external drive environments. They include diverse file types, sizes, and real-world data sources to ensure robust testing.