Digital Archive

Mar 24, 2020

Contents:

CHAPTER 1

CLI

This implements the Command Line Interface which enables the user to use the functionality implemented in the `digiarch` submodules. The CLI implements several commands with suboptions.

1.1 digiarch

Used for indexing, reporting on, and identifying files found in PATH.

```
digiarch [OPTIONS] PATH COMMAND1 [ARGS]... [COMMAND2 [ARGS]...]...
```

Options

--reindex
Whether to reindex the current directory.

Arguments

PATH
Required argument

1.1.1 checksum

Generate file checksums using xxHash.

```
digiarch checksum [OPTIONS]
```

1.1.2 dups

Check for file duplicates.

```
digiarch dups [OPTIONS]
```

1.1.3 group

Generate lists of files grouped per file extension.

```
digiarch group [OPTIONS]
```

1.1.4 identify

Identify files using siegfried.

```
digiarch identify [OPTIONS]
```

1.1.5 report

Generate reports on files and directory structure.

```
digiarch report [OPTIONS]
```

CHAPTER 2

Identify

2.1 Generate Checksums & Check for duplicates

This module implements checksum generation and duplicate detection.

check_collisions (*checksums: List[str]*) → Set[str]
Checks checksum collisions given a list of checksums as strings. Returns a set of collisions if any such are found.

Parameters **checksums** (*List[str]*) – List of checksums that must be checked for collisions.

Returns A set of colliding checksums. Empty if none are found.

Return type Set[str]

check_duplicates (*files: List[digiarch.internals.FileInfo], save_path: pathlib.Path*) → None
Generates a file with checksum collisions, indicating that duplicates are present.

Parameters

- **files** (*List[FileInfo]*) – Files for which duplicates should be checked.
- **save_path** (*Path*) – Path to which the checksum collision information should be saved.

checksum_worker (*file_info: digiarch.internals.FileInfo*) → digiarch.internals.FileInfo
Worker used when multiprocessing checksums of FileInfo objects.

Parameters **fileinfo** (*FileInfo*) – The FileInfo object that must be updated with a new checksum value.

Returns The FileInfo object with an updated checksum value.

Return type FileInfo

file_checksum (*file: pathlib.Path*) → str
Calculate the checksum of an input file using BLAKE2.

Parameters **file** (*Path*) – The file for which to calculate the checksum. Expects a *pathlib.Path* object.

Returns The hex checksum of the input file.

Return type str

generate_checksums (*files: List[digiarch.internals.FileInfo]*) → List[digiarch.internals.FileInfo]
Multiprocesses a list of FileInfo object in order to assign new checksums.

Parameters **files** (*List[FileInfo]*) – List of FileInfo objects that need checksums.

Returns The updated list of FileInfo objects.

Return type List[FileInfo]

2.2 Identify Files

Identify files using siegfried

identify (*files: List[digiarch.internals.FileInfo], path: pathlib.Path*) → List[digiarch.internals.FileInfo]
Identify all files in a list, and return the updated list.

Parameters **files** (*List[FileInfo]*) – Files to identify.

Returns Input files with updated Identification information.

Return type List[FileInfo]

sf_id (*path: pathlib.Path*) → Dict[pathlib.Path, digiarch.internals.Identification]
Identify files using siegfried and update FileInfo with obtained PUID, signature name, and warning if applicable.

Parameters **path** (*pathlib.Path*) – Path in which to identify files.

Returns Dictionary containing file path and associated identification information obtained from siegfried's stdout.

Return type Dict[Path, Identification]

Raises IdentificationError – If running siegfried or loading of the resulting JSON output fails, an IdentificationError is thrown.

update_file_info (*file_info: digiarch.internals.FileInfo, id_info: Dict[pathlib.Path, digiarch.internals.Identification]*) → digiarch.internals.FileInfo

2.3 Generate Reports

Reporting utilities for file discovery.

report_results (*files: List[digiarch.internals.FileInfo], save_path: pathlib.Path*) → None
Generates reports of explore_dir() results.

Parameters

- **files** (*List[FileInfo]*) – The files to report on.
- **save_path** (*str*) – The path in which to save the reports.

CHAPTER 3

Data & Utilities

3.1 Data

3.2 Path Utilities

Utilities for handling files, paths, etc.

explore_dir(*path: pathlib.Path*) → digiarch.internals.FileData

Finds files and empty directories in the given path, and collects them into a list of FileInfo objects.

Parameters **path** (*str*) – The path in which to find files.

Returns **empty_subs** – A list of empty subdirectory paths, if any such were found

Return type List[str]

CHAPTER 4

Exceptions

CHAPTER 5

Indices and tables

- genindex
- modindex
- search

Python Module Index

d

Index

Symbols

C

D

E

F

G

I

P

R

S

U