***ddeedduupp* v0.1a**

**User Manual**

January 1, 2020

David T. Ashley ([dashley@gmail.com](mailto:dashley@gmail.com))

Table of Contents

[1 License 3](#_Toc28684462)

[2 Introduction and Overview 3](#_Toc28684463)

[2.1 Brief Description of *ddeedduupp* Functionality 3](#_Toc28684464)

[3 Background Information 4](#_Toc28684465)

[3.1 Tree Nomenclature 4](#_Toc28684466)

[3.2 The SHA-512 Cryptographic Hash 5](#_Toc28684467)

[4 Detailed Description of *ddeedduupp* Functionality 5](#_Toc28684468)

[4.1 General Notes and Cautions 5](#_Toc28684469)

[4.2 Rules for Choosing Which Duplicate to Retain 6](#_Toc28684470)

[4.3 Invocation 6](#_Toc28684471)

[4.4 Command-Line Options 7](#_Toc28684472)

[4.5 Subcommand 8](#_Toc28684473)

[4.6 Detailed Description of the Base Directory 9](#_Toc28684474)

[4.7 Detailed Description of the Reference Directory, Reference Set, and Base Set 9](#_Toc28684475)

[4.8 Output, Error Termination, and Exit Code 9](#_Toc28684476)

[5 Invocation Example 10](#_Toc28684477)

[6 Technical Description and Internal Operation of *ddeedduupp* 10](#_Toc28684478)

[6.1 Heap Allocation 10](#_Toc28684479)

[6.2 RAM Data Structures 10](#_Toc28684480)

[6.3 Rebuilding *ddeedduupp* from Source Code 10](#_Toc28684481)

[6.4 Modifying *ddeedduupp* 10](#_Toc28684482)

[*6.5* Redistributing Modified *ddeedduupp* 10](#_Toc28684483)

[7 References 11](#_Toc28684484)

[8 Index 12](#_Toc28684485)

# License

*ddeedduupp* (source files, binaries, documentation, and all other files) is made available under the *MIT License*, reproduced below.

Copyright 2020 David T. Ashley  
  
Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:  
  
The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.  
  
THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

# Introduction and Overview

## Brief Description of *ddeedduupp* Functionality

*ddeedduupp* is a simple *Windows* and *\*nix* tool for identifying and selectively deleting duplicate files. It was originally devised by the author to identify and eliminate duplicate photos and documents.

The general characteristics of *ddeedduupp* are:

* On invocation, it calculates the SHA-512 cryptographic hash of every file in the set of files in which duplicates are to be identified.[[1]](#footnote-1)
* The SHA-512 cryptographic hash is used to quickly identify duplicates.[[2]](#footnote-2)
* *ddeedduupp* is written in ANSI C and should compile on virtually any platform. Full source code and a *Windows* binary are provided. For non-Windows systems, the end user will have to rebuild the program from source code.

The limitations and shortcomings of *ddeedduupp* are:

* The program performs a narrow range of de-duplication functions (one size likely does not fit all).
* On each invocation of the program, the SHA-512 hash of every file is calculated. This is potentially a very time-consuming operation.
* The program supports only ASCII characters in file names, and does not support Unicode. It is not known how the program will behave on operating systems where directory and file names may contain non-ASCII characters.
* *ddeedduupp* requires that the pool of files in which duplicates are to be identified and removed be staged in a single directory on a single disk volume.
* *ddeedduupp* is single-threaded. (A substantial performance increase during the initial SHA-512 calculation might be possible on some platforms if a multi-threaded model were adopted, but this has not been evaluated.)
* Unlike other more sophisticated de-duplication programs, *ddeedduupp* has no notion of files which are *nearly* identical. *ddeedduupp* will identify and remove only files which are *exactly* identical.
* The program forms a tree data structure in RAM to represent the directory and file hierarchy of the set of directories and files being examined for duplicates, as well as to contain the SHA-512 cryptographic hashes of files. Approximately 1 gigabyte of RAM is necessary for every 5 million files to be examined. There may be uncommon cases (a RAM-constrained machine combined with a large number of small files to be examined) that will cause a computer to run out of RAM.

# Background Information

## Tree Nomenclature

On invocation, the first step performed by *ddeedduupp* is to identify the files and directories in the base directory. Internally, *ddeedduupp* forms a tree data structure in RAM.

The notion of a tree to represent a computer file system is probably familiar to most computer users. Only directories (not files) may have children, and the children of directories may only be other directories. From *ddeedduupp*’s point of view, files are part of a directory (i.e. a file is *not* a child of its containing directory).



Figure : Example Tree (from Wikipedia article https://en.wikipedia.org/wiki/Tree\_structure)

Figure 1 (copied from [this](https://en.wikipedia.org/wiki/Tree_structure) Wikipedia article) is a prototypical tree to illustrate the nomenclature for the relationships between directories as used in this document and in the output from *ddeedduupp*.

* *Node*: Each directory is a node. *Encyclopaedia*, *Science*, *Culture*, *Art*, and *Craft* are nodes.
* *Child*, *Children*: Only a directory may have children; and the children, if any, are other directories. *Science* is a child of *Encyclopaedia*, and *Craft* is a child of *Culture*.
* *Parent*: The directory containing another directory. *Culture* is a parent of *Art*, and *Encyclopaedia* is a parent of *Science*.
* *Root*: A node with no parent. Only *Encyclopaedia* is the root.
* *Leaf*, *Leaf Node*: A node with no children.
* *Sibling*: A node with the same parent. *Art* and *Craft* are siblings.
* *Uncle*: A sibling of a node’s parent. *Science* is an uncle of *Art*.
* *Ancestor*: A node that can be reached by traveling up the tree. *Craft* has the ancestors *Culture* and *Encyclopaedia*.
* *Descendant*: A node that can be reached by traveling down the tree. *Encyclopaedia* has all other nodes as its descendants. *Culture* has *Art* and *Craft* as its descendants.

## The SHA-512 Cryptographic Hash

The SHA-512 cryptographic hash is part of the SHA-2 [] suite, and is an NIST standard. Numerous web pages exist describing the algorithm and providing reference code.

*ddeedduupp* uses the SHA-512 cryptographic hash to identify likely duplicate files. Although the SHA‑512 cryptographic hash is quite reliable in determining whether two files are identical, *ddeedduupp* compares files before deleting to be certain that the files are identical, and will not perform a deletion if two files have the same SHA-512 cryptographic hash but are in fact non-identical files.

How likely is it that two files that are not identical have the same SHA-512 cryptographic hash?

* No SHA‑512 hash collisions have been found to date.
* 2512 is approximately 10154. (For comparison, the number of atoms in the observable universe is estimated to be 1082, a much smaller number. It is *extremely* unlikely that two non-identical files would have the same SHA-512 hash; but *ddeedduupp* does protect for this possibility.)

# Detailed Description of *ddeedduupp* Functionality

## General Notes and Cautions

1. On \*nix systems, *ddeedduupp* follows both hard and soft links.
2. On Windows systems, *ddeedduupp* follows soft links.
3. *ddeedduupp* makes the assumption that it is the only program modifying the filesystem of the target directory. Any modification of the filesystem of the target directory while *ddeedduupp* is running may have unpredictable results, including loss of data.
4. *ddeedduupp* commands that remove duplicates are not undoable. The duplicate files are deleted permanently.
5. *ddeedduupp* will *not* remove the last file of a set of duplicates (such an operation must be done manually, using some method other than *ddeedduupp*).

## Rules for Choosing Which Duplicate to Retain

In contexts where *ddeedduupp* is to retain only one of a set of duplicates and where the choice of which duplicate to retain is not specified by other rules, the program chooses the duplicate to retain using these rules:

1. The duplicate highest in the directory tree is chosen.
2. If the previous rule results in a tie, the file in the directory whose components are first in sort order is chosen.
3. If the previous two rules result in a tie, the file name first in sort order is chosen.

Note that these rules tend to:

* Dispose of duplicates lower in the directory hierarchy.
* Dispose of duplicates whose ancestors are later in the sort order.

This typical effect of the rules is intentional. Duplication tends to involve entire directories of duplicate files, and retaining the duplicates that are higher in the directory tree and first in path sort order tends to create the result desired by most human users.

## Invocation

*ddeedduupp* is invoked from the command prompt (Windows) or from a shell (Unix) as follows:

ddeedduupp [options] subcommand base\_directory [reference\_directory]

The square brackets (“[options]”, “[reference\_directory]”) are used to indicate that an entity is optional. The square brackets may not be typed on the command line.

The options (§) are always optional.

The reference\_directory may or may not be required/allowed, depending on the subcommand (§).

The components of the command-line/shell invocation form above are explained briefly below; and in more detail in subsequent sections.

1. ddeedduupp  
   The program name.
2. [options]  
   Options that may modify how the program behaves (§).
3. subcommand  
   Describes the program’s major mode of operation (§).
4. base\_directory  
   The directory in which the program scans for duplicates (§).
5. [reference\_directory]  
   The reference directory, which must be an improper subdirectory[[3]](#footnote-3) of the base directory, potentially in conjunction with command-line options (§), serves to partition the file system objects into the base set and the reference set. This is described more fully in §.

## Command-Line Options

lists the options that can be used when invoking *ddeedduupp*. Not all options are applicable to all subcommands.

Table : ddeedduupp Command-Line Options

|  |  |  |
| --- | --- | --- |
| **Command Option Number** | **Command** | **Action** |
|  | *-dryrun* | Causes the command to be executed with no effect on the file system. This provides a preview of what the program would do if invoked without the *‑dryrun* option. |
|  | *-notouch\_rd* | Prohibits the program from deleting files within the reference directory itself. (However, this does not prohibit deletion of files in ancestors or descendants of the reference directory.) |
|  | *-notouch\_rd\_descendants* | Prohibits the program from deleting files within the descendants of the reference directory. (However, this does not prohibit deletion of files in the reference directory or its ancestors.) |
|  | *-refset\_include\_rd* | Causes the reference directory to be included in the current reference set.  This option is at this time never necessary, because the reference directory is always included in the reference set by default. |
|  | *-refset\_exclude\_rd* | Causes the reference directory to be excluded from the reference set. |
|  | *‑refset\_include\_rd\_descendants* | Causes the descendants of the reference directory to be included in the reference set.  This option is at this time never necessary, because the descendants of the reference directory are always included in the reference set by default. |
|  | *‑refset\_exclude\_rd\_descendants* | Causes the descendants of the reference directory to be excluded from the reference set. |
|  | *‑refset\_include\_rd\_ancestors* | Causes the ancestors of the reference directory to be included in the reference set. |
|  | *-refset\_exclude\_rd\_ancestors* | Causes the ancestors of the current reference directory to be excluded from the current reference set.  This option is at this time never necessary, because the ancestors of the reference directory are always excluded from the reference set by default. |
|  | *-refset\_include\_rd\_siblings* | Causes the siblings of the reference directory to be included in the reference set. |
|  | *-refset\_emptydir\_remove* | After duplicate files are removed, any empty directories (directories that contain no directories or files) are removed.  Directories are analyzed and removed from the bottom of the tree up, so empty directories that contain only empty directories will also be removed. |

## Subcommand

describes the *ddeedduup* subcommands.

Table : Subcommands

|  |  |  |  |
| --- | --- | --- | --- |
| **Subcommand Number** | **Subcommand** | ***reference\_directory* Argument Supported** | **Description** |
|  | *noop* | Yes | Performs a full analysis, but performs no deduplication operations. This command can be used for two purposes:   * To get information about the base directory and reference directory (and descendants), including information about number and identity of duplicates. * With the *‑refset\_emptydir\_remove* option, to delete empty directories but perform no other operations. |
|  | *dd\_bs* | No | All duplicate files within the base set are removed. |
|  | *dd\_rs\_pri* | Yes | All files in the base set that are duplicates of file(s) in the reference set are deleted.  No file within the reference set is deleted. |
|  | *dd\_rs\_sec* | Yes | All files in the reference set that are duplicates of files(s) in the base set are deleted.  No file within the base set is deleted. |

## Detailed Description of the Base Directory

The base directory (see item 4, §4.3) is the directory in which *ddeedduupp* operates to identify and remove duplicates. *ddeedduupp* will not open, read, or delete a file or directory not in the base directory and its descendants.

When *ddeedduupp* is initially invoked, the program calculates the SHA-512 cryptographic hash of every file in the base directory and its descendants.

## Detailed Description of the Reference Directory, Reference Set, and Base Set

The reference directory (see item , §4.3) must be specified as an improper descendant of the base directory.

In specifying the base directory and reference directory on the command line:

* No referential nomenclature for paths (“.”, “..”) is allowed by *ddeedduupp*.
* *ddeedduupp* must be able to verify that the reference directory is an improper descendant of the base directory. This means that the specified string representing the base directory must appear at the start of the specified reference directory.

*ddeedduupp* conceptually calculates the reference set and base set as follows:

* The *reference set* is the reference directory, combined with all its descendants, but adjusted by any command-line options that modify the reference set.
* The *base set* is the base directory and all its descendants, with the elements of the reference set removed.

Note that:

* The reference set and base set have no intersection.
* The union of the reference set and the base set is the base directory and all its descendants.
* It is possible for the reference set or base set to be empty.

## Output, Error Termination, and Exit Code

The program writes informative and error output to *stdout* and error output to *stderr*.

It is safe to redirect *stdout* to a file but leave *stderr* directed to the console: any error messages will be written to both *stdout* and *stderr*, so that the lack of console output would mean that the program completed successfully.

Nearly all errors and suspicious situations are treated as an error and will result in program termination. If the program completes without errors, the exit code will be 0; an error termination gives an exit code of 4.

# Invocation Example

# Technical Description and Internal Operation of *ddeedduupp*

## Heap Allocation

## RAM Data Structures

## Rebuilding *ddeedduupp* from Source Code

## Modifying *ddeedduupp*

## Redistributing Modified *ddeedduupp*

# References

[1] SHA-2, <https://en.wikipedia.org/wiki/SHA-2>

# Index

ancestor

defined, 6

base directory, 5, 10

defined, 10

base set, 8

defined, 10

child

defined, 6

*dd\_bs* (subcommand), 9

*dd\_rs\_sec* (subcommand), 10

descendant

defined, 6

improper subdirectory, 7

leaf

defined, 6

license, 4

MIT license, 4

node

defined, 6

*noop* (subcommand), 9

parent

defined, 6

reference directory, 10

reference set, 8

defined, 10

root

defined, 6

SHA-512 cryptographic hash function, 6

sibling

defined, 6

tree, 5

uncle

defined, 6

Unicode, 5

1. Calculating the SHA-512 cryptographic hash of every file allows *ddeedduupp* to identify duplicates very quickly once the initial calculation is done. [↑](#footnote-ref-1)
2. Please see §. [↑](#footnote-ref-2)
3. *Improper subdirectory* is used here in the same sense as *improper subset* in mathematics—it is allowable for the base directory and reference directory to be the same directory. However, the reference directory may not encompass file system objects not also in the base directory, i.e. it must not be above or distinct from the base directory. This is described more fully in §4.7. [↑](#footnote-ref-3)