Babbacombe Computers Ltd
File DeDuplicator

File Deduplicator (Dupe) is a Windows 3 program which searches hard and floppy disks for files which have the same name, are the same size, or have identical contents. It allows you to restrict the search to files whose names match a "Mask" (like a DOS wildcard name, eg *.* is all files and *.txt is all files with the extension "txt"), and to files whose sizes are in a certain range. In addition you can select which directory (or directories) to search, and carry out a search across several devices (so, for example, you could tell it to look for all bitmap and backup files which are in your c:\windows directory and on a floppy) or across a network.

Having found some duplicate files File Deduplicator allows you to delete any of the files that it has found. It doesn't delete them until you go back to the main screen, and checks you really do want to delete them, so don't be afraid to experiment.

History

V1.0 (11 July 1991)
Original version

V1.1 (15 July 1991)
Able to work in Real Mode
Includes Cancel Button during searches
Fully Multitasking during searches
Reports if unable to compare 2 files
Allows choice between confirmation of all deletions, or individual confirmation for each file
Changed search by name to search by full name and added new search by name

V1.2 (2 August 1991)
Produce Report of Search Results to File or Printer.
Display Details of selected files.
Search Results window may be iconised.
Yields control to Windows while comparing file contents, improving multitasking.
Improved keyboard control (NB this has caused the directory selection method to change. You must now double click on a directory to select it).
Includes optional virus detection check on loading.
Includes optional 3D appearance dialog boxes (NB three_d.dll must be in the path even if this is not used).
Includes Online Help.
No longer uses WIN.INI. It now uses an initialisation file specific to Babbacombe Computers products. If the registration information is entered in an older version then it will need to be reentered when installing this version or later.

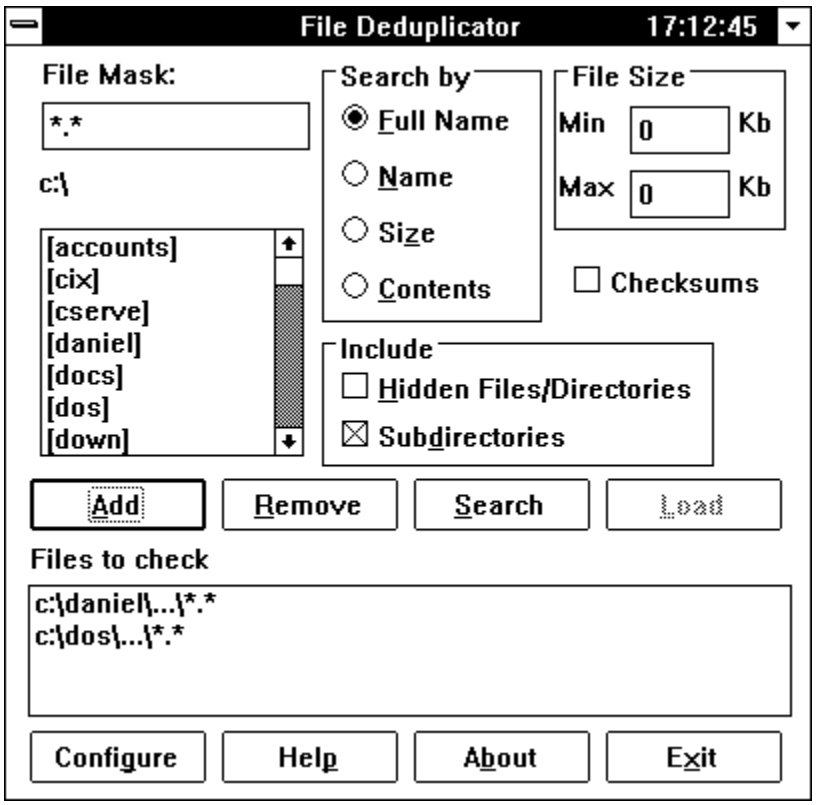
V1.3 (22 August 1991)
Gives more information on the progress of searches
Allows the windows to be minimised during searches
Warns if an attempt is made to search too many files
Allows multitasking to be switched off during searches
The "Constructing File List" and "Retrieving File Info" Phases have been amalgamated
Reports the number of (possible) duplicate files found
Utilises checksumming to speed up searches by content

V1.4 (5 January 1992)
Limit on number of files in a search removed
Enhanced selection of directories to include in a search
Can Save Search Results to disk, and reload them
Apparent hangup problem at end of searches fixed
Bug which prevented deletion of files on some systems fixed
Can delete read-only files after confirmation
Can be iconised while printing
Displays "Page Number" on Search Results Window
Includes Configuration Screen for setting defaults
Allows Browse of Search Results by Name or Path
Can selectively undelete files
Includes delete flag in prints
Allows prints of files marked for deletion only
Includes a file viewing facility (uses an external viewer)
Will run without three_d.dll being present
Focus is on Multitask switch during Searches (prevents accidental cancel from spacebar)

V1.4a (10 January 1992)
Switch to control failure messages
Failure Log facility
System Share Violation and Cannot Read from Drive message boxes disabled
Enter key "double click" restored on directory list box
Wider scroll bar in Results Window
Auto Selection of results

The Main Screen

The Main Screen of File Deduplicator looks like this.



The **Files to check** box near the bottom of the screen will list the directories and files which you want to search for. At first this is blank, and the Remove and Search buttons are greyed out so that you cannot use them.

You add to the list by selecting the file mask which you want to look for (eg *.* for all files, a*.bmp for all bitmaps starting with 'a') and the directory or directories which you want to search. The current directory appears directly below the file mask. You change it by clicking on entries in the list box below it. This is similar to selecting a file in Notepad, except that you select only a directory. When you have selected the directory you want to search click the add button and the directory/filemask will appear in the **Files to check** box. You will notice that they are separated by '\...\'. This means that all subdirectories of the directory you have selected will also be searched. If you switch off the **Subdirectories** checkbox and click the **Add** button again then a new entry will appear without the '\...\'. If you click **Add** while the **Hidden Files/Directories** box is checked then '[H]' will be put after the entry in the **Files to check** box. This means that hidden files and subdirectories will be included in that search (NB System files are never included in any search).

Alternatively, rather than adding the current directory to the list you can select one or more of its subdirectories in the directory list box. If any of the entries in the list box are selected when you press the **Add** button then they will be added to the **Files to Check** box instead of the current directory. This makes it easier when you want to exclude some subdirectories from a search.

Don't worry if you include some files twice in the search, eg by selecting 'c:\...*.bmp' and 'c:\windows*.*'. The program will not list all of the bitmap files in c:\windows as duplicates of themselves (although the search may take slightly longer, though not much).

Once you have an entry in the **Files to check** box the **Remove** and **Search** buttons become usable. If you select an entry (or several entries) in the box and click the **Remove** button it (or they) will be removed from the search.

In the top right hand corner of the window is a **File Size** box where you can enter 2 numbers, Min and Max. These specify the minimum and maximum sizes of the files which you want to have included in the search. If Max is zero (or less than Min) then there will be no limit on the size of files included. Unlike the **Hidden** and **Subdirectories** options these apply to all of the entries in the **Files to check** box at the time you click the **Search** button, rather than being set separately for each entry.

Next to the **File Size** box is the **Search By** box. This allow you to select which files are considered to be possible duplicates. If the **Full Name** button is on then the program will consider all files with the same name to be possible duplicates, and will list them by name (eg c:\windows\party.bmp and a:\picts\party.bmp will be listed together). The **Name** button is similar but the program will consider all files with the same name up to the '.' to be possible duplicates (this is handy for looking for .bak etc files). If the **Size** button is on then the program will instead consider all files of the same size as possible duplicates, and will list them by size. If the **Contents** button is on then the contents of the files will be compared and only files which are actual duplicates will be listed together. It may seem odd to give the other 3 options when this one is available, but the reason is that searching by contents can take a lot longer than searching the other ways. If you do a search by name or size you can still check whether the files which are listed really are duplicates, and this may be quicker.

The **Load** Button allows saved Results files to be reloaded into Dupe. See the section on the **Save As** menu item in the Search Results window below.

The **About** button gives some information about the program. It also allows you to register your copy so that the nagging doesn't happen (but you have to pay for it first).

The **Help** Button runs calls up the Windows 3 Help system to provide Help for Dupe.

The **Configure** button brings up the configuration window. See below for the explanation.

The **Checksums** box will only have any effect during a search by Content. If checked then Dupe will calculate full check sums for each selected file before comparing their contents. Otherwise it will calculate partial checksums "on the fly". Generally, it is expected that performance will be better with the option off, but in some circumstances it can be dramatically improved by calculating full checksums before starting. See the technical notes below for more details on this option.

When you are ready to begin the search click on the **Search** button. Your machine will chunter away to itself for a while and then tell you the results of the search. If a box comes up saying "No files were selected" then it means that the search turned up nothing to report (not that there are no files in those directories). If it has found possible, or genuine, duplicates then the "Search Results" window will appear.

While the search is being done this small window will appear which tells you the current phase of the search, how many files it has found and, once it starts comparing the contents of files, how far through the list it has worked.

Note that the percentage figure given when Comparing File Contents refers to how far along the list the lowest file being compared is. It does not always give an accurate estimate of how much longer the search will take as in testing this proved to be impossible. If the number increments very much more slowly during one portion of the "Comparing File Contents" phase then it may be worth setting the Checksums option for the search. See the Technical Notes below for details.

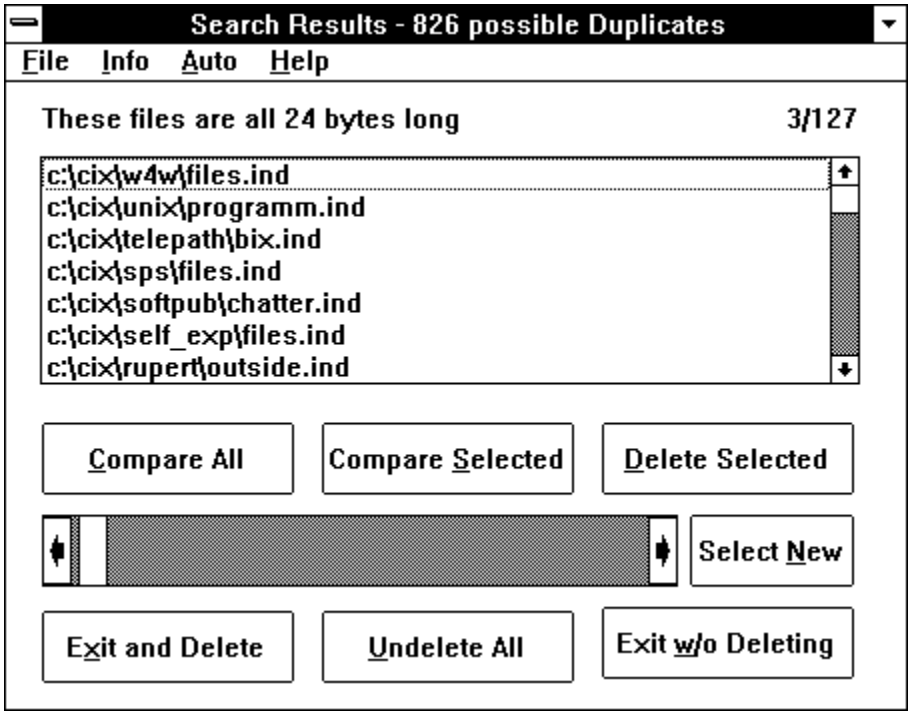
It also contains a **Cancel** button and a **Multitask** Checkbox. If you click on the button then the search will be cancelled. If the screen is cluttered during a search then clicking on the minimize button on this window will iconise it and the main window. If you wish to have this window on the screen during a search, but would like to clear the main window off the screen, then you may do this by iconising this window and restoring it by double clicking on the icon's title bar, rather than the icon itself. You can restore the main window by iconising again and double clicking on the icon. This same technique will work with the Search Results window, as well. The main window is always restored to the screen at the end of a search or when leaving the Search Results window.

While the Multitask box is checked (as it always is at the start of a search) the program will check as often as possible to see whether other applications are operating and will effectively give priority to them as far as possible, so you can click on any other window (except for the Main Window of File Deduplicator), or launch a new application, and it should work almost as fast as usual. You may notice that it slows down slightly, particularly when File Deduplicator is comparing files on a floppy, but that can't be helped. If you click on the Multitask box and remove the checkmark then Multitasking will effectively cease to operate, which should speed up the search, although you will probably not notice a significant effect unless the search takes a very long time. While multitasking is off, Dupe will still occasionally allow messages, such as mouse clicks, to be received, otherwise it would not be possible to cancel a search once multitasking was off, but it allows it relatively rarely. As a result, if you click on the Cancel or Multitask buttons subsequently they may take some considerable time to respond. If you wish to switch multitasking back on take care not to click on the checkbox twice, thinking that the first click may not have been received, as you will only succeed in switching it back off immediately.

When comparing files by content you may receive a number of message boxes informing you that Dupe was unable to compare two files. This usually occurs because one or other of the files is locked (in use by another application), but may also happen if a file is deleted or a floppy changed during the search. If you are running Dupe across a network then you may receive a very large number of these messages. You can disable them, and optionally log them to a disk file, by using setting the relevant Configuration options (see below). Dupe will consider any files it fails to compare to be non-duplicates.

The Search Results Window

This window looks like this:



At the top of the window there will be a message telling you about the files listed. If you searched by full name it will tell you that these files all have the same full name. If you searched by name it will tell you that these files all have the same name. These two options will also tell you the size of each file. If you searched by size it will tell you the size of the files. If you searched by content it will tell you that all of the files listed are identical. If there is more than one set of possible duplicates (eg after a search by size there may be a list of files of 0 bytes and another of files of 100 bytes) then the scroll bar allows you to move along the pages. If there is only one page of possible duplicates then the scroll bar will not appear. The far right hand side of the message shows the current "page" number an d the total number of pages of (possible) duplicates.

The **Compare All** button will tell you whether the contents of all of the files in this list are identical to all of the others. If 2 are but a third is different it will simply tell you that they are different.

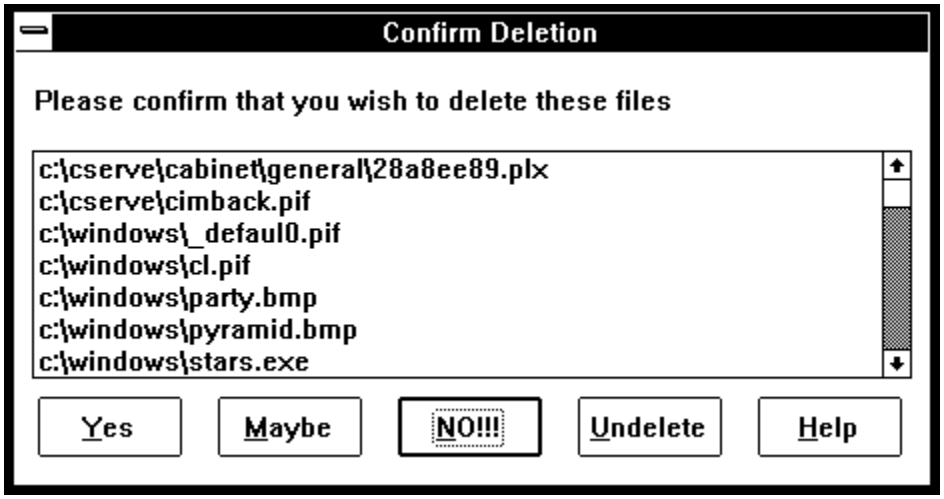
If you select 2 or more files from the list the **Compare Selected** button will compare them. Like the **Compare All** button it will only tell you whether all of the files you have selected are identical. The **Compare All** and **Compare Selected** buttons are greyed out after a search by content, because you already know that all the files in one list are identical.

The **Delete Selected** button will flag the selected files for deletion. The program doesn't actually delete them yet, though. What it does is remove them from the list of files it is displaying. If you delete all of the files on one page then that page will still be displayed but will be empty.

The **Undelete All** button will remove the flags against all the files you have deleted (not just the ones on the page you are looking at), and redisplay the file names in the box.

The **Exit w/o Delete** Button simply returns to the Main Screen.

The **Exit and Delete** button will display a window showing you the list of files which you have deleted and will check that you do want to delete them. This window has 4 buttons, marked **Yes**, **Maybe**, **No** and **Undelete**. If you press **Yes** then the files will be deleted and you will return to the Main Screen. If you press **Maybe** then the program will ask for confirmation before deleting each file. If you press **No** then you will be returned to the Search results Window. If you press **Undelete** after selecting some files from the list then all of the selected files will be reinstated on their respective pages, no longer marked for deletion. If you undelete all of the files then you will return to the Search Results window. The Confirm Deletion screen looks like this:



The program will check each file before attempting to delete it to ensure that it is not a read-only file. If it is then it will ask for confirmation before deleting it.

The **Exit and Delete** button and the **Undelete All** button are greyed out until you delete something.

To the right of the scroll bar is a button called either **Select Old** or **Select New**. If the search was by Full Name or Name it will be **Select Old**, if it was by Size or Contents it will be **Select New**. This button allows you to automatically select all but the newest file in the list, if it is **Select Old**, or all but the oldest if it is **Select New**. This is useful in cases where a lot of duplicates have been found and you know that only one is required (the latest in a search by name, and the earliest in a search by Contents). The default setting (Old or New) can be changed by selecting the required item from the Auto Menu.

The Search Results window has a File Menu (with the entries Save As, Print, Print Deleted, and Printer Setup), an Info Menu (with the entries Display Details, View File, Browse by Name and Browse by Path), and an Auto Menu (with the entries Select Old and Select New).

Selecting **Save As** brings up the standard Save As dialog which allows you to save the current state of the Search Results Window to a disk file. As well as the list of

(possible) duplicates Dupe will also save the list of deleted files, and the current page. You can reload the saved file by pressing the **Load** button on the Main Window and selecting the file previously saved. This effectively puts the program back in exactly the same state as it was in when the file was saved. The filename defaults to "*.dpe", but any other file name, and directory, may be set as the default by changing the "Save:" setting in the Configuration window.

The **Print** item will allow you to print a report of the files found in the search. It will ask whether you want to print to a file or to the printer, and whether you want a summary or full report. If you send the report to a file then it will prompt you for a file name. A summary report simply lists the possible duplicate files found in the search. Each set of files is separated by a dashed line. A full report also gives details of the search criteria used and a line giving the details of each file (the same details as the Display Selected item below). Deleted files are marked on these listings with an asterisk '*'. A report sent to the printer will be paginated according to the current printer setup and will put a heading and page number on each page. A report sent to a file will not include any headings or pagination. When printing to a printer a dialog box will be displayed which will allow the printing to be cancelled.

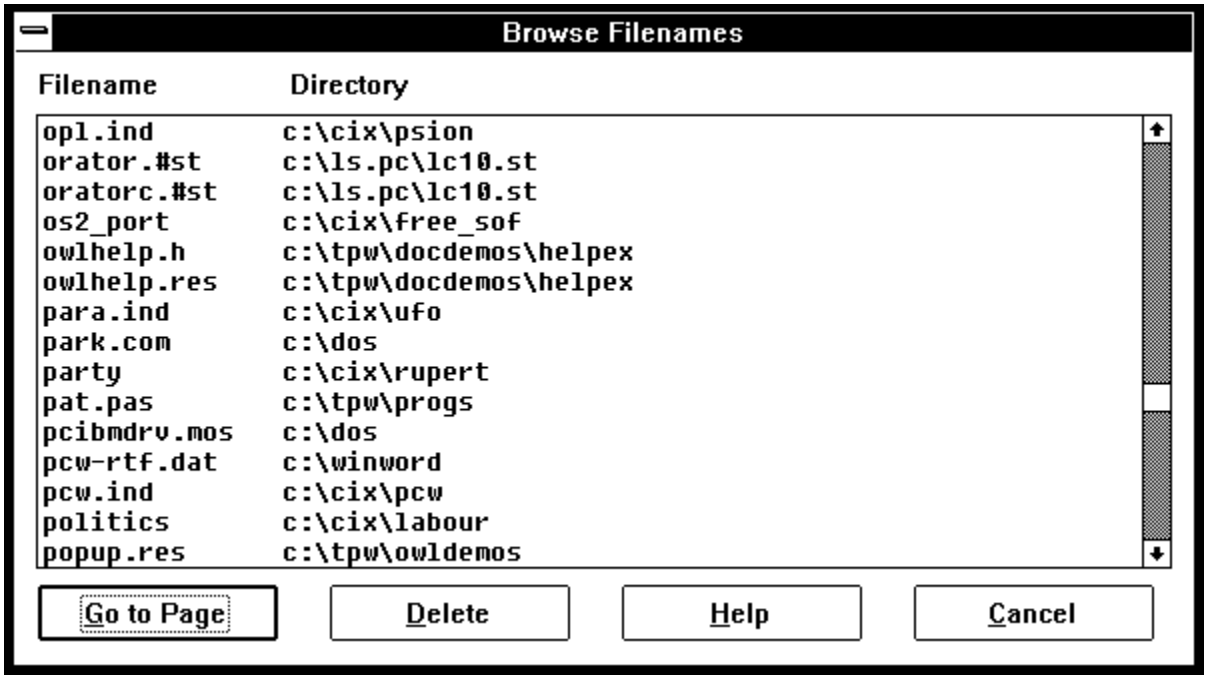
The **Print Deleted** item will produce listings similar to that from the **Print** item, but containing only deleted files. The files in these listings are not specifically marked as deleted, and they are not split into sets.

The **Printer Setup** item will enable you to change the setup for the current printer. The information you can set depends on which printer you have. You cannot change to a different printer by using this option, and the changes made are temporary. To change the current printer, or make permanent changes to the set up, you should use the Control Panel.

If you select one or more files from a list and select **Display Selected** then the program will display a box for each selected file (up to 5) listing its last modification date and time, its size and its file attributes.

If you select a file and select the **View File** item a file viewer program will be started up to enable you to see the contents of the file. No viewer is supplied with Dupe, by default it uses the Notepad accessory supplied with Windows. This is not ideal for a number of reasons. It is limited as to the size of files it can load, and can only really display text files (and even then a number of common text formats are not supported by it). It is recommended that you configure Dupe (using the Configuration window explained below) to use a shareware or freeware Windows file viewer. Any viewer which can accept the filename as a parameter can be used. If you do not have a suitable Windows viewer then Vernon Buerg's LIST.COM can be used (although this is a DOS program you can, if you use enhanced mode, set up a PIF file to run it in a window). The **View by Association** item is similar, but if a program has been defined for the file using the Associate option in the File Manager then that will be run instead of the default viewer.

The **Browse by Name** and **Browse by Path** items will display a dialog which lists all of the (possible) duplicate files which are not marked for deletion. The dialog may take a few seconds to appear, as the files have to be resorted first. The former shows the filename followed by the directory of the file and is sorted by the filename. The latter shows the full pathname and is sorted by this. For example, the **Browse by Name** dialog looks like this:



If you select a file and press the **Go to Page** button (or double click on a file) then you will return to the Search Results Window page which includes that file. If you select one or more files and press the **Delete** button then those files will be marked for deletion, and removed from the Browse List. Pressing the **Cancel** button will return you to the Search Results window. Note that pressing the **Cancel** button does not undelete any files marked for deletion from this dialog.

The Virus Detection Option

File Deduplicator contains code which allows it to carry out a simple check to see if its executable file has been modified by a virus. This is not a rigourous virus check and should not be used instead of specialised virus detection programs. However, it is possible that it could cause a virus to be detected earlier than it otherwise would, and this could limit the damage caused.

By default, the virus checking facility is switched off. It can be enabled from the Configuration window (see below).

With virus checking switched on the program will take several seconds longer to start up, depending on the speed of your machine (about 5 seconds longer on mine).

Any modifications to the executable file will cause the virus check to be triggered (if it is switched on).

3d Dialog Boxes

Dupe contains code which enables it to use the facilities provided by a Dynamic Link Library called three_d.dll. This library allows programs which use it to display their Dialog Boxes using a sculpted 3D effect. Dupe does not insist that you use the library, and can display its Dialogs with the standard Windows look. This can be set on or off from within the Configuration window as explained below. Dupe will operate even if the library is not available but will then, of course, always use the normal Windows

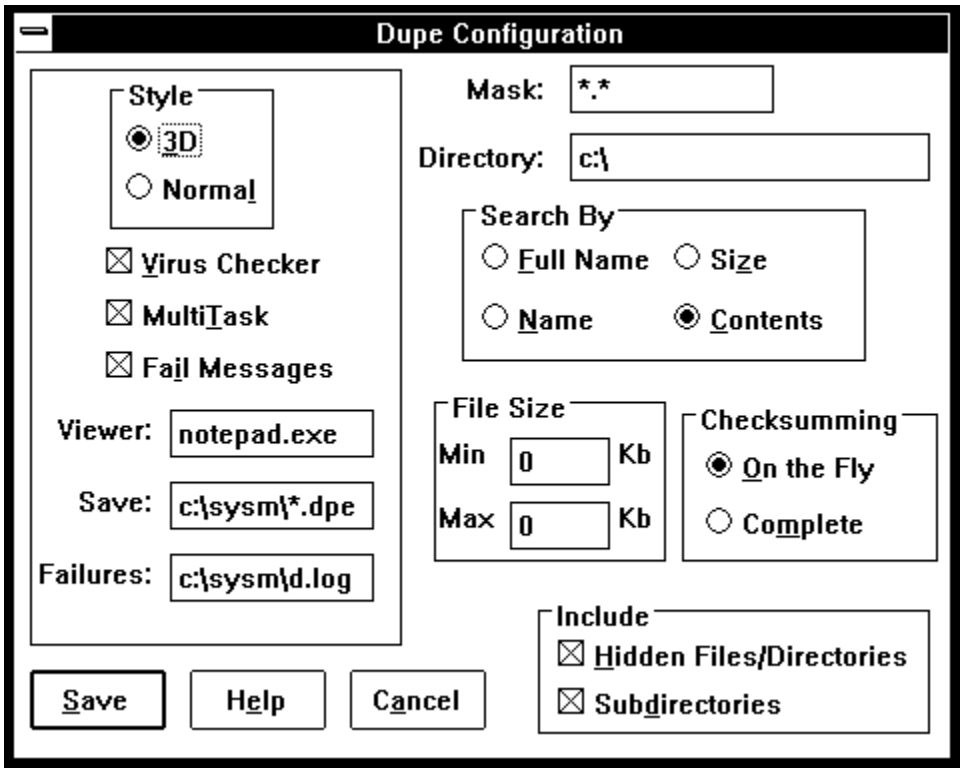
style.

The library file three_d.dll can be placed in any directory on your path (ie included in the DOS PATH environment variable). Your Windows directory, or the directory where you place utilities, is suitable.

With 3d styling on the virus detector will still use normal Windows styling to report errors. With luck, you will never see this.

The Configuration Window

This is reached by pressing the **Configure** button in the Main window, and looks like this:



The Style box is used to set whether Dupe will use the 3D style or the Normal Windows style. It does not effect until the next time Dupe is run. If three_d.dll is not available then it will use the normal style regardless of this setting.

The Virus Checker switch determines whether the Virus check will be run when Dupe is started up. It does not take effect until the next time Dupe is run.

The Multitask switch is used to set the default setting of the Multitask switch during a search. When a search is started the multitask switch in the Cancel dialog will have the same setting as this switch, but can still be changed during the search. This switch takes effect immediately.

The Fail Messages switch allows you to define whether Dupe will inform you of failures to compare files, due to Share violations for example, by using a message box. If it is off then you will not be informed of any failures, in which case you may want to log them instead (see the Failures box below). This switch takes effect immediately.

The Viewer box allows you to specify a file viewer to be used in place of notepad.exe when viewing files you may wish to delete. This option takes effect immediately.

The Save box allows you to specify the default file mask for saved results files, and takes effect immediately.

The Failures box allows you to specify the name of a file to which failures during the search will be logged. If it is empty then the failures will not be logged. If the file already exists then new messages are appended to it. If a file is specified and the Fail Messages switch (see above) is off then you will be informed that failures have occurred at the end of the search. If there are any problems writing the log then you will be informed that logging failed at the end of the search. This option takes effect immediately.

The remaining options set the default values which will appear in the main window when Dupe is started up. For example, if you always carry out searches by Content then set the Content switch in the Search By Box, or if you usually carry out searches starting in the \USER directory on disk D: then set the value in the Directory box to d:\ user and set the Subdirectories switch on. All of these options take effect the next time Dupe is run.

If you find that the directory box and current path box are empty when you start Dupe then the default path you have specified is invalid. You should change it and restart Dupe.

When you have finished configuring Dupe press the **Save** button to store the new settings. Pressing the **Cancel** button will leave the settings as they were.

Technical Notes

Search size limit
Until Version 1.4 Dupe was unable to carry out a search on more than 16,380 files. This limit has now been removed.

Search Methods
It has proved to be effectively impossible to give an accurate estimate of how long File Deduplicator will take to carry out a search by Content. The documentation for early versions contained an estimate of around 15 minutes for 70 MBytes of files, based on tests run on the development machine. However, some users reported that searches on disks containing around 100 MBytes of files could take around 2 hours. Subsequently it was discovered that one directory on the development machine, accounting for less than 20% of the file space, took up 75% of the time for the search.

Investigation of the precise reasons for this led to the development of two very much faster algorithms for comparing the contents of files, which is an example of why we would like you to contact us with any problems which you encounter with the program.

This explanation of how the program operates is intended to help you choose which of the two algorithms, selected by the Checksums box in the main window, you should be using to carry out searches, given the types of files you have on your disks. If this is too heavy going then our advice is that the default setting (Checksums off) will probably be quicker, but that if you feel a search is taking too long then it may be worth trying it again with this option on.

Whichever type of search is being carried out Dupe first examines the directories listed in the search list and stores the path names of specified files together with file details, such as size. This list is maintained in such a way that a particular path name is only entered into the list once. At the end of the "Constructing File List" phase this list is resorted into the order required by the search type: by file name for Full Name and Name searches, and by size for Size and Content searches. The list is then

examined in a single pass, the "Eliminating non duplicates" phase, and entries which have no duplicates, according to the data in the list, removed. In searches by Content at this stage having no duplicates means only that a file has a unique size, in other words so far this search is identical to one by Size.

At this point searches by Full Name, Name and Size are complete, and it is only necessary to insert markers to identify each group of duplicates before the results can be displayed.

For a search by Content, however, it is now necessary to compare the files of each size to determine which of these have identical contents. For each set of files of a particular size, the program starts at the bottom of the list and compares the first file with each of the others in turn. It marks each file which duplicates the first and, in effect, removes them from the set of files still needing comparison. If none of them duplicate the first one it discards it. If there are still files of this size in the list then it repeats the process, until either all but one have been discarded, or all of the remaining files are marked as being duplicates. A comparison is, of course, abandoned as soon as 1 byte is discovered to be different in both files.

In version of Dupe prior to 1.3 this was the only method used to compare files. For any set of files which were the same size but were not identical the time taken to examine them increased exponentially against the number of files (in fact, as the square of the number of files). So if comparing n files of the same size took t seconds, then comparing 2n files could take up to $4t^2$ seconds. It is clear, then, that the time taken for a search depended not on the number of files specified in the initial search, nor on the total size of the files searched, but on whether there were a large number of files of the same size which contained a large number of unique files (in the case where the files all duplicate the behaviour was linear but this, of course, is rare).

The comparison algorithm used from version 1.3 makes this behaviour almost linear in most cases, so that the search times depend more on the total number of files remaining after the "Eliminating non duplicates" phase, and doubling the number of files will usually roughly double the search time.

During the first run through comparing the contents of a set of files of the same size, the program always reads at least the first 4Kbyte block (or less if the files are shorter, of course) of each of the files in the list. In version 1.3 it takes advantage of this to compute a checksum of the contents of the first block of each file. During subsequent passes through the list it does not compare the contents of files which have differing checksums. It is uncommon for two blocks with different contents to have the same checksum, so the program is usually able to reject two non-duplicates without having to read them again. Where the checksums are the same the contents must, of course, be compared as before.

This strategy will usually speed up searches dramatically. However, because checksums are only calculated for the first 4K of each file, it is possible to find a situation where it confers little advantage. If the search includes a number of large files of the same size which differ only after the first block the behaviour of the search will be exponential, as before. To cater for this situation the search contains an option, switched on using the Checksums box, to calculate complete Checksums for all of the files whose contents are to be compared before starting the comparisons. Generally, it is anticipated that this calculation will take slightly longer than carrying out the comparisons using the on-the-fly checksumming method detailed above, which is why the option is off by default. If you notice that the percentage figure displayed during the "Comparing File Contents" phase is incrementing very slowly during a portion of the search, then it may be worth attempting it using the full checksums option. It is extremely unlikely, in all practical situations, that any search

carried out using this option will exhibit exponential increases in search times, but if anyone complains that this is still happening there are further devices which can be implemented which should cure it.

Acknowledgements

The Printer access routines are used by permission of OptiCom Inc, who can be reached via CompuServe as 71250,71.

The virus detection code is based on a program written by Nick Wallbridge and published in Personal Computer World (Aug 1991).

Three_d.dll is used by permission of Ray Donahue, 365 Mather Street, Unit 125, Hamden, CT 06514. If you wish to develop applications using the 3d styling in Turbo Pascal for Windows then you may wish to obtain a TPW Unit developed by Babbacombe Computers Ltd which simplifies development using the DLL, and which also provides the 3d styling for message boxes. Note that fees charged for the use of this Unit by Babbacombe Computers are in addition to those required by Ray Donahue to allow distribution of the DLL.

The Help screens were generated using Xantippe, from IRIS Media Systems, who can be reached via CompuServe as 76547,706.

Ordering

File Deduplicator V1.4a is a shareware program. You may use it for up to 21 days to decide whether it is useful to you. If you wish to continue using it after that then you must pay for it. The Registration fee is £15 sterling or $25 US. You can send cheques (in sterling or dollars) or cash to:

Babbacombe Computers Ltd
397 Meanwood Road
Leeds
West Yorkshire
UK
LS7 2LL

We can be contacted by email as tprinn@cix.compulink.co.uk or 100016,2726 on CompuServe or by telephone on 44 532 459673 (that is, 0532 459673 within the UK). Write or call for details of site licenses.

In return for registering we will tell you how to personalise your copy of the program and make it skip the initial nagging, and send you updates, bug fixes etc. (of Version 1) of the program if you require them (any updates will be released onto Cix and CompuServe as they are produced).

You may freely distribute unmodified, unregistered copies of this program to friends and colleagues, and upload it to Bulletin Boards, provided that you accept no payment for them, and that this documentation is included with them. Shareware vendors and user groups must contact Babbacombe Computers before including the program in any catalogues or distributions.

Order Form for File Deduplicator V1.4a

Send, together with a cheque for £15 or $25, to:
Babbacombe Computers Ltd
397 Meanwood Road
Leeds
West Yorkshire
UK
LS7 2LL

Name: _____

Address: _____

_____

_____

_____

Email: _____

If you wish to receive a copy of the program and updates on disk then tick the disk size. 3½ "___ 5¼"___
If you do not tick one then we will email (or snail mail) you as updates are released so that you can download the new version.

Name to appear in About Window (this is used to generate your registration number):

_____