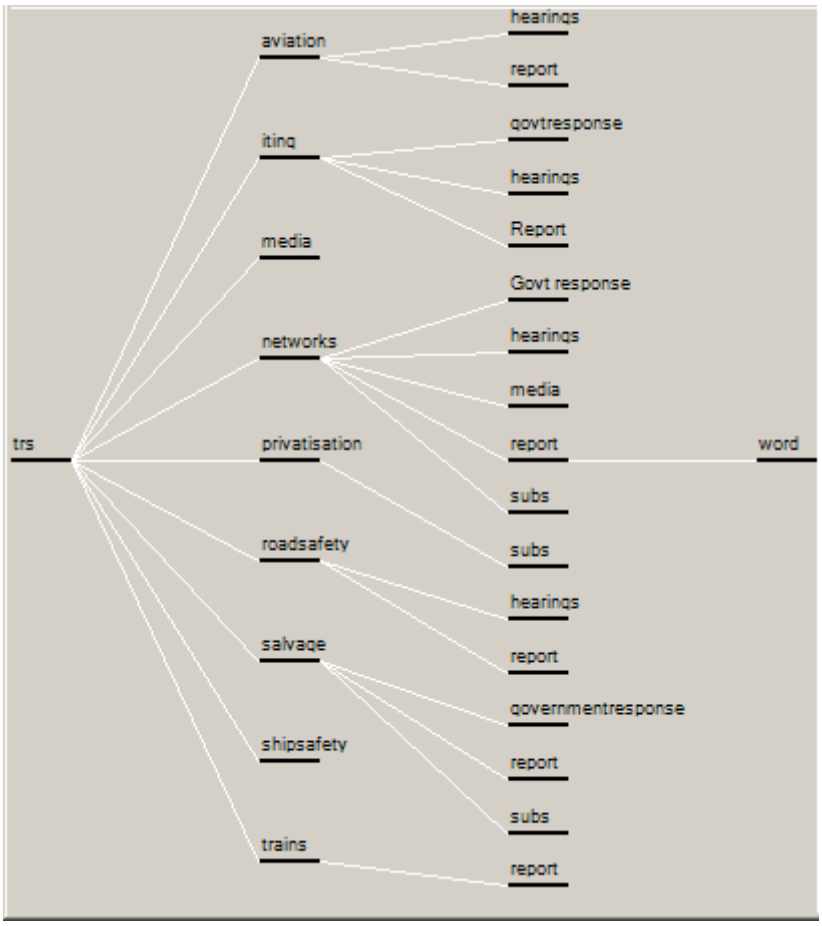
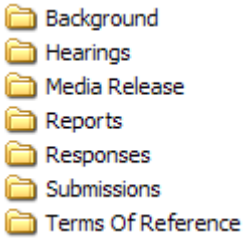
Filesystem Reorganisation Case Study

Data and Desired Structure

A small collection of 787 files spread over 47 folders in four levels contains documents relating to the activities of the Australian Parliament House of Representatives Standing Committee on Transport and Regional Services from the Australian Parliament House web site. The document are grouped into nine areas of Committee activity corresponding to the 2nd level folders. The documents comprise submissions, background papers, hearing transcripts, media releases, terms of references and responses to submissions. The documents are mostly PDF files with some in Word format. The folder structure is shown below:

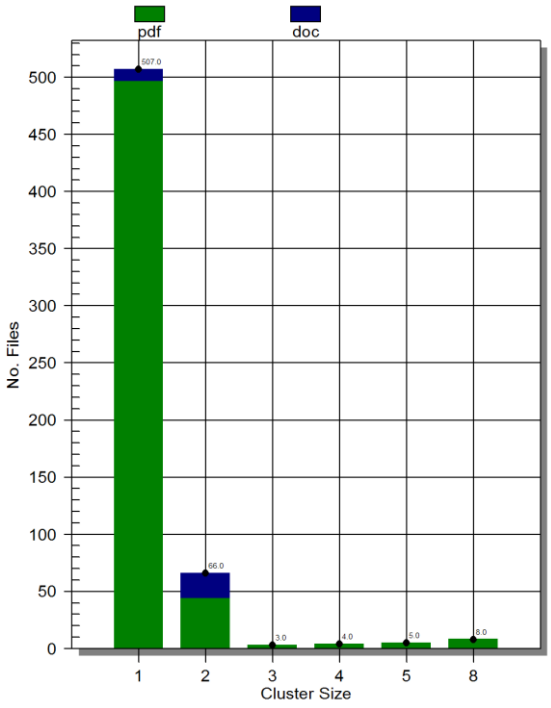


The desired destination tree is a single level structure grouped by document type as shown below:



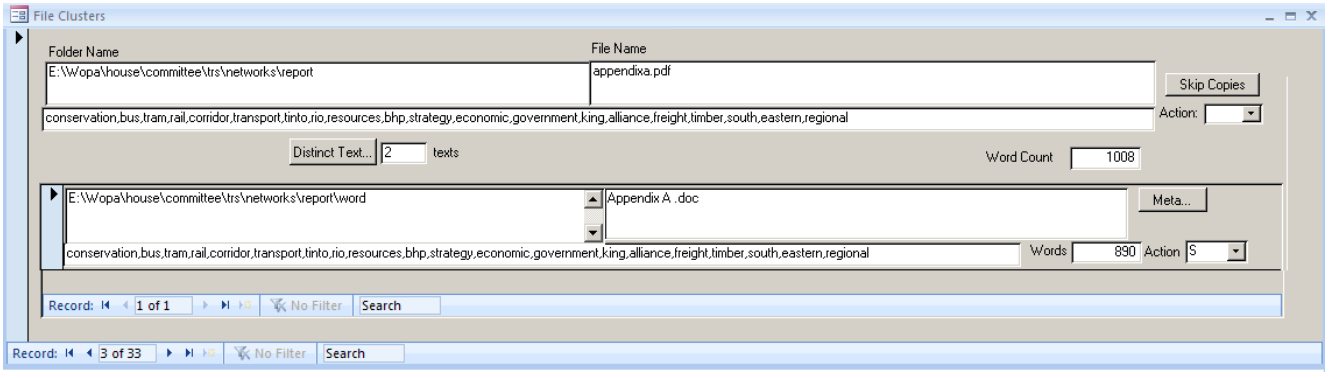
Duplicates and Early File Versions

The first step is to identify groups of files with similar or identical text content, and flag copies or early file versions for skipping during the reorganisation process, using the interface shown in Figure 1.



• Figure 1 Text Near and Exact Match Spectrum

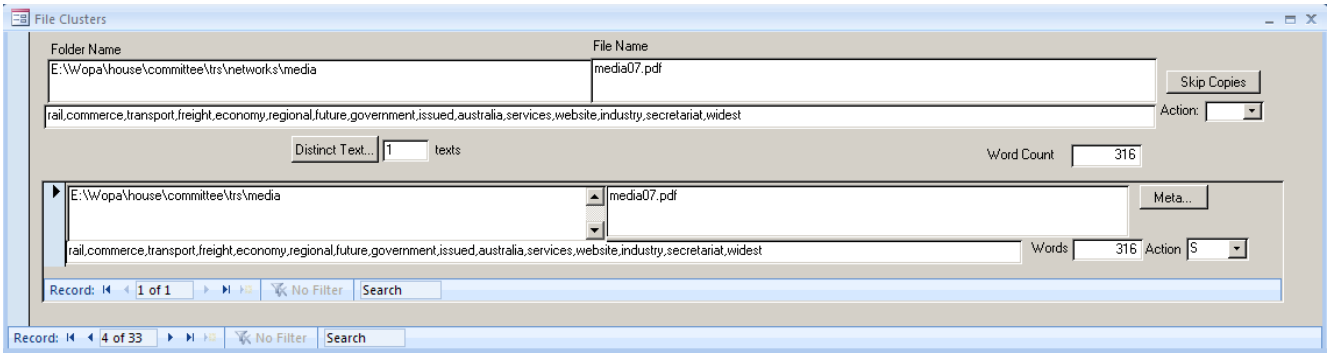
Each cluster of 2 or more files can be viewed and the most recently modified file selected, as shown in Figure 2 for a cluster of two similar but not identical files which constitute versions of the same file, with the older file in Word format and flagged for skipping



• Figure 2 Example of a near-matching cluster

File text content can be viewed to determined if the near-matching status is correct and the files can be considered as different versions of the same document.

Figure 3 shows an exact matching cluster where the file names are the same but they are in different folders. The older version is skipped here as well.

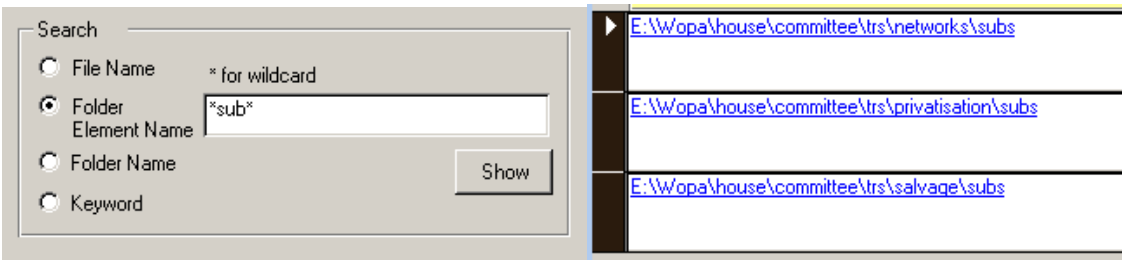


• Figure 3 Example of an exact-matching cluster

Once all 37 clusters have been examined and duplicates flagged for skipping (taking about 10 minutes), the destination folders for files can be set.

Mapping to Destination Folders

In many cases, the name of a folder indicates the destination folder (or container) required for the files in the folder and subfolders. For example, folder names containing the text ‘report’, ‘hearing’, ‘response’, ‘media’ and ‘sub’ will contain files belonging in the Reports, Hearings, Response, Media Releases and Submissions containers respectively. Files can be assigned to the appropriate container using the search facility to find folders containing the desired string:



• Figure 4 Result of searching for folders containing the string ‘sub’ in their name

All files in these folders and their subfolders can then be displayed and assigned to appropriate container. However, only 503 of the 787 files are included in folders meeting these criteria. The remaining files are contained in the level 2 folders with the names of the nine areas of activity. A listing of files in these folders indicates that files with a particular function indicated by the file name appear in these folders, as shown below:

1	E:\Wopa\house\committee\trs\roadsafety\tor.pdf
2	E:\Wopa\house\committee\trs\privatisation\tor.pdf
3	E:\Wopa\house\committee\trs\networks\tor.pdf
4	E:\Wopa\house\committee\trs\media\tnmed03.pdf
5	E:\Wopa\house\committee\trs\trains\titor.pdf
6	E:\Wopa\house\committee\trs\trains\timed01.pdf
7	E:\Wopa\house\committee\trs\aviation\sub99.pdf
8	E:\Wopa\house\committee\trs\aviation\sub98.pdf
9	E:\Wopa\house\committee\trs\aviation\sub97.pdf

• Figure 5 Part of list of files in level 2 folders

These files can be opened from within Drive Analyser and it can be determined that files containing the text ‘sub’ in their file name are submissions. These files can then be selected and then assigned to the

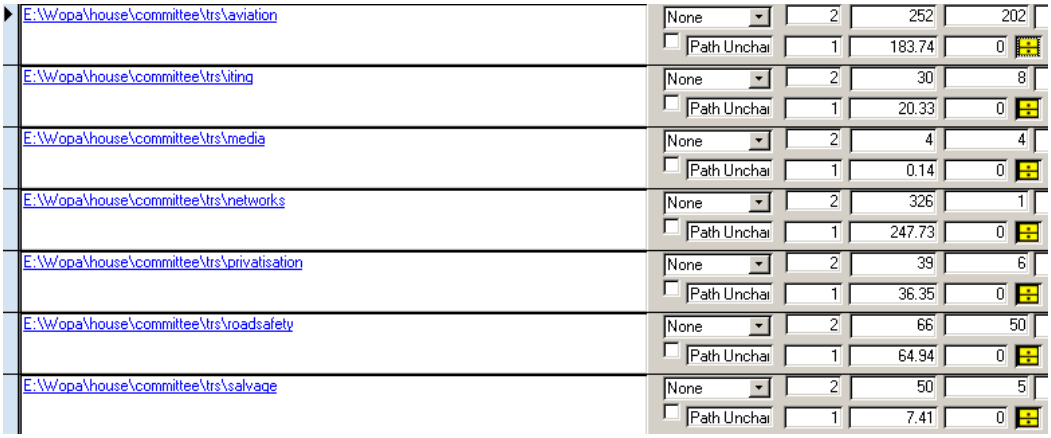
'Submissions' container. The same process can be used to assign files whose name contains 'tor' to the Terms of Reference container, and files whose name contains 'med' to the Media container. The small number of files which do not contain any of these strings and are in folders which have not already been assigned to containers can be assigned individually. The files are listed below:

E:\Wopa\house\committee\trs\Warehouse to Wharf Report.pdf
E:\Wopa\house\committee\trs\shipsafety\back.pdf
E:\Wopa\house\committee\trs\salvage\dispaper.doc
E:\Wopa\house\committee\trs\salvage\dispaper.pdf
E:\Wopa\house\committee\trs\privatisation\prpaper.doc
E:\Wopa\house\committee\trs\shipsafety\chap1.pdf
E:\Wopa\house\committee\trs\shipsafety\chap2.pdf
E:\Wopa\house\committee\trs\shipsafety\chap3.pdf
E:\Wopa\house\committee\trs\shipsafety\chap4.pdf
E:\Wopa\house\committee\trs\shipsafety\chap5.pdf
E:\Wopa\house\committee\trs\shipsafety\front.pdf
E:\Wopa\house\committee\trs\shipsafety\ship.pdf
E:\Wopa\house\committee\trs\shipsafety\chap6.pdf
E:\Wopa\house\committee\trs\privatisation\prpaper.pdf

• Figure 6 Files not containing strings 'sub','tor', 'med' which are not in folders already assigned to containers

Defining Destination File Names via Title Metadata

Now that all files have been assigned to containers, they can be assigned names indicating their provenance which would come from the folder name in the original folder tree.The first step is to show all the folders at the level from which provenance needs to be assigned as shown below:



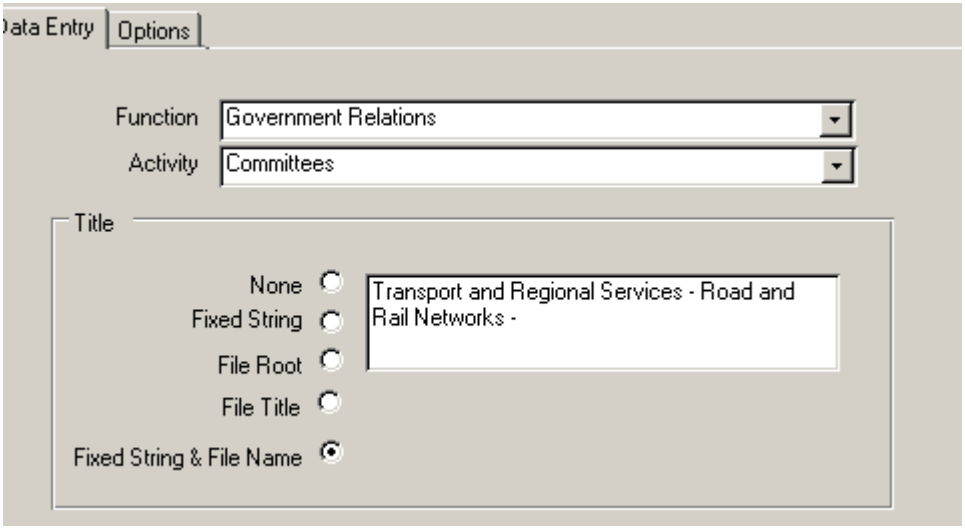
• Figure 7 Folders at level 2 used for defining file name to reflect provenance

Clicking on the yellow cabinet icon shows all files below the folder.

Part of the display of files below the aviation folder is shown below:

- Figure 8 Part of list of files below the Aviation folder

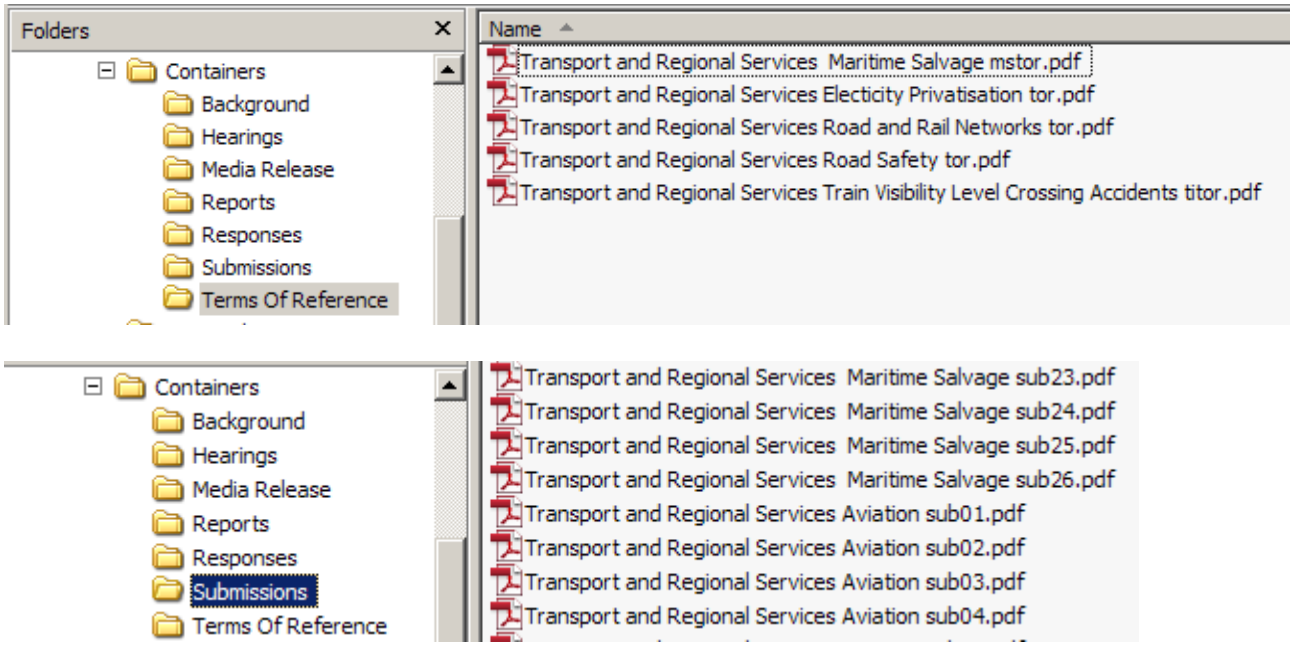
Files selected from this display can then be assigned Title metadata comprising a fixed string and the file name, as shown below. Function/Activity metadata from a thesaurus such as Keyword AAA can be applied at this point.



- Figure 9 Assignment of title metadata indicating file provenance and Function/Activity metadata which may be used in Document Management System to assign sentences.

Final Results

Once Titles have been assigned for all the folders defining areas as shown in Figure 7, all the files can be copied to containers using the Title metadata as a file name to give results as shown below:



- Figure 10 Results of file re-organisation

The files are now grouped in containers reflecting the type of document with the area to which it relates recorded as part of the file name. Processing for this re-organisation took 20 minutes.