Locating files on computer disks

Jonathan Jermey

Over the last 20 years file storage capacity on personal computer systems has grown from the equivalent of a desk drawer to that of a medium-sized public library, and the trend is continuing. File retrieval methods have been developed on an ad hoc basis which has generally failed to keep up with this growth, but there are some encouraging and interesting developments in this area.

When I began using a personal computer (PC) in 1986, it had a total storage capacity of 720 Kb, represented by two floppy disk drives. The largest number of files that I had access to at any one time was about 20, and the total number I had in my collection at its peak was perhaps 200. None of them were sound, font or image files. I had lots of computer difficulties and problems back then, but finding files was not one of them.

From there I graduated to a 40-Mb hard disk and then to a 300-Mb disk and 1.44-Mb floppies. Shortly after this CD-ROMs put in an appearance and eventually became cheap enough to install on a PC. Our current PC has 13Gb of storage capacity and is currently half-full with about 75,000 files. Add about 75Gb of CD-ROM storage, and we end up with something like a million files available to us. In 15 years PCs have become mass storage and retrieval systems. And the Internet adds literally billions of files to those available.

Meanwhile, the growth in storage capacity is continuing with DVD and other technologies. Imagine indexers equipped with 'lifetime disks' – single portable storage elements about the size of a credit card, containing every single index they have ever done, including all the revisions; every single invoice they have ever written; every single book they have indexed; every email they have sent or received; and details of every person they have ever contacted, every financial transaction they have made, every government or commercial document they have ever been notified of or sent, perhaps even every book that they have read, every piece of music they have ever listened to and every video or TV program they have seen. At current rates of storage capacity growth, this will be attainable in another 15 years or so.

Unfortunately, systems for organizing and retrieving this information have not kept up with its growth. The concepts of information organized into files, and files organized into directories, still dominate computer storage systems as they did in the 1960s and 1970s. Indexing and cataloguing techniques have a part to play in providing access to this material, but currently it is only a small one. In what follows, I discuss some old and new techniques available to PC users for organizing and retrieving files on their own systems. Although the examples relate to Windows 98, similar methods are available to Mac and Linux users.

Filenames

Each file on a computer system has a name, usually made up of two parts, the name proper and an *extension* indicating the type of file. Within any one directory, all files must have different names, but files in different directories may have the same name, whether or not they have the same content. Since the extension is used to indicate file type, this leaves the name proper to identify the file. Prior to Windows 95, file names were restricted to eight characters and could not contain spaces; although these restrictions have now been lifted (the limit is now 246 characters), many systems of naming still stick to the old rules for various reasons. Among these are:

- compatibility with older systems, or because their files were created under older systems;
- automated naming systems, where the name is given by software rather by than a human decision;
- dependence on other programs that still require specific short names for their ancillary files;
- concern that users may be disconcerted by the sudden appearance of very long file names.

Thus the Galaxy of ClipArt collection from RomTech, for instance, contains a picture file called 'KKCHLD02.WMF' which could be named more appropriately: 'Childish drawing of a person with a green face and long wavy hair'. There is no reason in principle why much more information could not be incorporated into a file name in this way, but it is unlikely to happen in the near future.

Location

The most widely used indicator of file contents is their location in the hierarchical storage system of folders and subfolders on the hard disk. In earlier days it was possible to maintain a rigid structure on the hard disk where C:\WORD\DOCUMENTS\LETTERS\EUROPE\SPAIN, for instance, would be a directory containing copies of all letters written in Word that had been sent to Spain. Changes in Windows 95 and 98 have made this rigidity less practical, but it is still possible to organize a hard disk or CD-ROM into a set of directories containing files of a particular type.

Setting up a workable directory structure requires classification skills, however, and a quick survey of materials on offer reveals how rare these are. Looking at the Galaxy of

ClipArt again, we find a variety of different classification systems overlapping on just one CD-ROM: directories are established on the basis of:

- target audience (e.g. KIDSTUFF)
- topic (e.g. RELIGION, NEWAGE, SCIFI)
- function (e.g. SIGNS)
- chromacity (e.g. BW, COLOR)
- location (e.g. HOME, for pictures of items found at home)
- source (e.g. GOSPELS)

There are also categories that defy analysis – STORY and ID, for instance. The division of directories into subdirectories is arbitrary, at best, and someone attempting to find a file through its location is likely to make many trials and errors in the process. The method used by Microsoft for arranging its clipart on disk is no better: clipart files are scattered across the system apparently at random, and directory names are used indiscriminately to indicate source program (e.g. PUBLISHER), target audience (e.g. CORPBAS = corporate basics?), and suggested usage (e.g. WEBART).

There is scope here for an indexer or cataloguer to set up a clear-cut directory structure for storing files, particularly since the arrival of shortcuts in Windows now makes it possible to provide 'cross-references' from one directory to files stored in another. Similar considerations apply to collections of recorded sound effects and videos.

The situation for music files is a little better, since these usually have clear-cut antecedents: a performer and/or composer, and a CD from which they come. Thus a MUSIC directory might contain subdirectories for BEETHOVEN and BRAHMS and under BEETHOVEN subdirectories for SYMPHONY No. 1, etc. Most new music CDs have an ID number embedded in the data on disk, and modern music copying programs can use this to look up details of a CD on the Internet while it is being copied. These details are used to create a directory structure for storing the resulting files, down to one directory per CD: any additional breakdown (e.g. into multiple works on one CD) must be done by the user.

Windows Find File

The Find File system is built into Windows 95/Windows 2000. It brings up a dialog box that allows a disk or a network to be searched for files matching certain criteria. These include file name, extension, size, dates of creation or modification, and – for text files only – text contained in the file. This latter suffers from all the well-known limitations of text-based searching, plus a few of its own; for instance, a search for 'fish' brings up a selection of system files that cannot be read or opened by the user, as well as Word and text files that may be of value. It can be used with non-text files provided they have embedded text fields (see below).

Thumbnail collections of picture files

A rudimentary type of indexing for picture files only can be done through the provision of a set of 'thumbnails' – small versions of each image that appear on the screen 20 or 30 at a time. These may be generated and then stored for use again (Paint Shop Pro does this with its 'Browse' command) or simply produced as required for a particular directory or disk of graphics files. The user can then select one of the thumbnails to open the associated file. Obviously thumbnails are effective only when there is a relatively small collection of graphics files to choose from; to browse through the thumbnails of a million files would take several weeks.

External databases

More sophisticated collections of material use external databases. These record the name and location of the file and then allow the user to add information about the picture, such as keywords, categories and sources. A good example is the Microsoft Clip Gallery program – especially after its recent facelift – which allows users to search not only for graphics but also for sound files and videos. New material is added to the Gallery either by installing new Microsoft programs (e.g. FrontPage) or by downloading new clips from the extensive collection on the Web, where the same type of keyword or category search is available. When new material is added to the disk the associated database records are incorporated into the Gallery database, so that users can search for files by keyword and/or category (one file can be placed in several categories). In addition, the Clip Gallery uses thumbnails to display search results, making it possible for the user to 'point and click' on their final choice. The user can add their own files to the Clip Gallery and (if they want to) add categories and keywords of their own to both new and old files.

Because the information in the Gallery links to the file via its name and location, if these change then the connection is lost and the file becomes an 'orphan'. This makes it difficult to copy or transfer material after it has been indexed. Microsoft has developed a format for compressing and packing picture files along with the associated database records, and it uses this for downloading clip art from its website, but the format is not available for the general public to use.

Some music file playback programs, like Real Jukebox, have carried external databases to a logical extreme, with dozens of fields available for each track: a music enthusiast can include the names of session musicians, a scanned image of the album cover, front and back, as well as the lyrics of the song and much more.

More serious applications for document databases include *document control.* For example, in some legal offices the creation and management of each electronic copy of every document can be tracked through a database back to its original creator. Security features can include restricting access to certain documents and automatic archiving or deletion of documents after a certain date. These systems are expensive to set up and cumbersome to maintain, but they allow for precise control and retrieval across a company-wide document database.

Embedded text fields

For most purposes it is best to have the file indexing information embedded into the file in some way, so that copies of

files retain this information. This has at last been recognized by software developers and new file formats are usually created with at least one text field available to be filled in when a new file of this type is created. At one end of the scale there are GIF files, with one text field available (Description); at the other the new Paint Shop Pro format (.PSP) offers fields for Image Title, Artist Name, Copyright and Description. Any of these description fields could in theory be used for embedded indexing. Text in these fields can then be found through the Windows Find File system already mentioned.

Particular applications like Word and Excel have their own way of 'tagging' files with properties that can then be previewed or searched for within the program. For Word these are visible through File/Properties and include Title, Subject, Author, Manager, Company, Category, Keywords and Comments. Other 'customized' properties are also available. Once these properties have been set up for a document collection the File/Open/Tools dialog box then allows the user to define the criteria for locating the file or files that should be opened. The properties 'travel with' the file so that if a Word file is sent by email, for instance, it retains its property information. Web authors will see parallels here with the metadata that can be associated with Web pages.

Embedded text fields are also used for audio files: MP3 format files, for instance, have four embedded fields for Track Name, Artist Name, Album Name and Genre. Thus when an MP3 file is copied or moved it carries a record of its origins with it.

Conclusion

The skills of indexers (and cataloguers) is gradually being recognized in the preparation of software for distribution: most large software companies now employ indexers to index their help files, for instance. As yet they are seldom called in after the event to organize material in a company document collection, for instance, or a clipart library; but these are also areas where we can make an important contribution. A little knowledge of file formats and storage and retrieval methods will give us credibility in discussing and negotiating for these jobs. Three areas can be identified in which indexers could be needed:

- Many large image collections are currently almost useless because of the difficulty of retrieving the images that are wanted. This problem can only worsen as collections become exponentially larger over time. Even cursory indexing could improve retrieval rates many times over and add value to the collection.
- Large companies with document control systems are currently relying largely on incidental information – who wrote the document, who typed it – and text-based retrieval to manage their collections. As these grow, the market for content-based semantic analysis of these documents will increase, allowing for retrieval on the basis of topics discussed or concepts raised.
- Collectors of music find it notoriously difficult to tell in advance whether they will like a particular piece or performer. Apart from 'genre' – a useful but nebulous MP3 category including such overlapping options as 'Show Tunes', 'Satire' and 'Pranks' – the only criterion for selection is usually a rough indication of similarity to some performer or piece that the user already likes. If a fully fledged system for indexing musical characteristics could be developed, then a ready-made target market would be available for it.

Apart from job-related considerations, though, there are other reasons to keep abreast of the systems and methods developed for computer information retrieval, especially when they are in use on our own desktop. After all, for many of us, our main competition comes from automated systems – and knowing your enemy is an important part of survival.

Jon Jermey is an indexer and computer trainer with a special interest in applying electronic methods to indexes, especially for ebooks. He has been Web Manager for the Australian Society of Indexers since 1998. Email: webindexing@optusnet.com.au