Software Archeology: An Experiment

Randall Neff

November 26, 2007

An example shown several Software Preservation Group meetings ago included adding a PowerPoint presentation to a proposed software archive. I was concerned about keeping software in proprietary, undocumented formats. I was also concerned that the metadata did not include a version number for the format, and did not include a list of required fonts. I think that software in proprietary formats should be immediately converted to public standard formats and the conversions stored along with the original. So a PowerPoint file would be converted to: rtf and a png image sequence; or pdf; or odt. In all cases except the image sequence, the fonts must be stored as well.

With this concern, I decided to try an experiment to see if lots of different file formats would be a problem with future understanding of software and files. I would pretend that my home PC (running Windows XP) would be 'found' and software archaeologists would want to analyze my disk contents. The machine has an 80 gigabyte drive split into two partitions and a 160 gigabyte drive split into two partitions. A file suffix is defined as characters after the last period in the file name.

Data from November 26, 2007

1. Files: 237,613
2. Bytes: 141 gigabytes
3. Files with no suffix: 11,512
4. Different file suffixes: **1542**
5. End Points of Suffix Occurrence Histogram

   421 suffixes occurred 1 time

   209 suffixes occurred 2 times

   ...

   1 suffix occurred 14245 times (.tif)

   1 suffix occurred 30334 times (.html)

6. Top ten suffix occurrences

Suffix	**Public**	**Occurrences**	**Total bytes**
html	Yes (text)	30334	489 Mbytes
tif	yes	14245	40 Gbytes

gif	Yes (patent expired)	12314	147 Mbytes
jpg	Yes	11840	2.8 Gbytes
<none>		11512	1.2 Gbytes
mp3	Sort of	8887	31 Gbytes
png	Yes	8198	176 Mbytes
tmd	No (RealPlayer)	7598	18 Mbytes
bmp	Yes	6953	163 Mbytes
h	Yes (text file)	6360	45 Mbytes

Conclusion

I guessed there would be about 200 to 300 different file suffixes on my disks. I was very surprised, there are **1542** different suffixes. The **11512** files with **no** suffix will be difficult to identify what the file contains and its format.

Software archeology is / will be a very difficult task even with contemporary PCs due to the enormous number of different file formats, and the requirement to document them all.

Randall.