



Quality Assurance in PDF - Business critical files may not be legible in the future

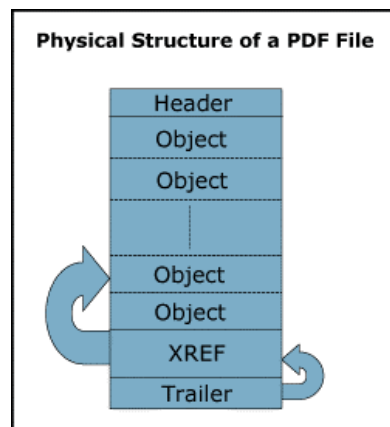
If you bake a cake, you start by checking the "best before" date on the cake mix, smelling the milk to ensure it's still good, and looking at the eggs when cracking them open. If any of the ingredients are bad or have reached their expiry date, you won't use them. In comparison, how many companies check the PDF documents they receive from external (or other internal) sources before entering them into business processes, where the cost of failure is considerably higher?

PDF is the preferred processing and archiving format for millions of business documents that have to be retained and reproducible for years. But it is alarming how few users are aware of the potential quality problems with PDF or analyze the quality of their PDF documents. PDF files that are created and processed in your daily business can contain corruptions that allow the documents to be viewed and appended today, but may hinder or wholly prevent their reproducibility in the future.

Quality Assurance in PDF

- Hidden corruption may make PDF files illegible in the future
- Standard applications can unknowingly be corrupting your PDF files
- Improper testing methods may result in corrupt PDF files being accepted
- PDF files can be analyzed against corruption
- Analysis tools can be easily integrated into business processes

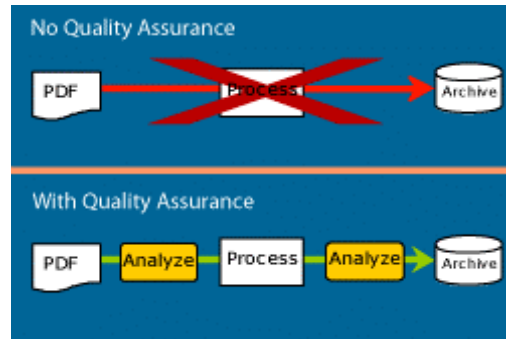
It may astound you, but basic operations like checking PDF into Visual SourceSafe or sending PDF by FTP may systematically produce corrupt PDF files. If VSS or the FTP applications are not setup to recognize binary files, the PDF files may be treated as text files and non-printing characters are automatically removed.



There are also endless possible inconsistencies with the semantics of imbedded files (fonts, Java script, XML's) and object attributes. These corruptions can be caused by creation, manipulation, or conversion processes. Another common cause is a file being truncated when it is transmitted. The physical structure of a PDF file (see picture) is quite different from it's logical document structure. First the header is read, which identifies the file as PDF, then the trailer. The trailer points to the cross-reference table, which then points to the objects containing pages, fonts etc. If the end of the PDF file is truncated, the trailer is incomplete and the process breaks down before the document can be read.

It is possible to view some corrupt PDF files with a PDF reader. Adobe® Acrobat for example repairs certain errors "on-the-fly" to make the PDF files viewable. However, it is optimized to compensate for corruption, and not to correct it. The future legibility of the PDF files is not guaranteed with this process. For this reason, using a specific version of Acrobat to verify the quality of PDF files does not guarantee that they are free from corruption.

The logical approach to guarantee the future legibility of PDF files is to properly analyze the files before they are entered into a business process. Corrupt files could be immediately identified and repaired or replaced. Once the business process (which could include a number of PDF manipulation and conversion functions) is completed, the output can again be analyzed to ensure that it is still valid.



Is this analysis really necessary? Let's put the question differently. Take for example older financial statements that were archived in PDF format. If you cannot quickly reproduce those statements when the tax auditor visits, how much effort will it cost you to reconstruct them?

Despite the necessity, there are relatively few comprehensive analysis tools available for PDF documents. One tool that can help here is the "3-Heights™ PDF Repair Tool" from pdf-tools.com. It was developed for internal quality assurance, i.e. to test and confirm the quality of the PDF documents that their own tools were creating and processing, and is now commercially available for both Windows and a variety of Unix platforms. The tool analyzes and repairs PDF files, and can recover information out of irreparable PDF files.

Another possibility is to use a PDF subset compliance check tool like "Colour Chameleon" from Grafikhuset, which cleans incoming Postscript before it hits Acrobat Distiller.

Integrating an analysis of PDF files into business processes can be quite easy, and could pay huge dividends in the long run. Conducting an analysis of current PDF archives is recommended to determine whether or not you already have been subjected to corrupt PDF files. It will also help you to identify possible sources of corruption and correct them.