

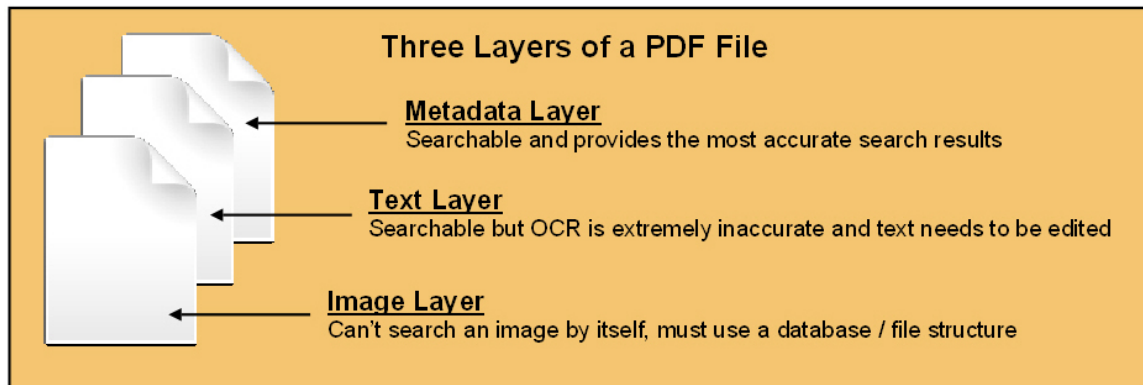
White Paper

Three Layers of A PDF File

**By: Randy Van Ittersum and Erin Spalding CDIA+
www.disusa.com**

How a Search Engine Finds a PDF File

For search purposes, visualize that a PDF file has three layers. This graphic illustrates the three layer concept and is essential to understanding how different document imaging systems search and retrieve electronic documents.



The first layer is the image layer

The first layer is a picture image of the paper document and can't be searched. There is no data that can be used in the search process other than the file name. Document imaging systems that use this layer must input the key search words into a database and hyperlink it to the image. Without the hyperlink connection, you cannot find your documents other than by file name.

Advantages: accurate searches

Disadvantages: high maintenance cost, high exit cost, and all of the inherent problems of a database

Recommended: only for large organizations that have an inside database IT person

The second layer is the text layer

From a scanned image you need to create text in order to have data that can be searched upon. This is done through a process called Optical Character Recognition or OCR. Keep in mind that OCR is not even close to perfect and is subject to errors that must be corrected in order to accurately search a document. For example, if the OCR process is 90% accurate it means that one in every ten characters is wrong. Studies conclude that when using everyday documents you will achieve about a 74% accuracy rate.

You can use an index to find documents using this layer but once you reach a large number of documents, this method becomes self defeating. The reason is systems that rely on this method, start to return high numbers of inaccurate search results. A search might return 1,000 results where there are only 10 documents that actually satisfy your search criteria. Most people don't have the patience to open 1,000 documents to find the information they need.

Advantages: uses an index and can be maintained by someone with basic computer skills

Disadvantages: inaccuracy of OCR and inaccurate search results

Recommended: desktop systems or organizations that have less than 15,000 documents

The third layer is the metadata layer

This layer is searchable by search engines and produces very accurate search results. Our system embeds the key search words that you assign to the document into the metadata layer. By embedding the key search words into this layer, we have in effect made the documents portable. Everywhere the document goes, the key search words go with it. This enables a search engine to easily find a document even after it has been moved from one file folder to another.

Advantages: accurate search results, uses an index, can be maintained by someone with basic computer skills, scaleable, low maintenance costs, and low exit costs

Disadvantages: none

Recommended: any size organization

Conclusion

We are of the opinion that every organization will derive significant benefits from a document imaging system. The next step is choosing the best system for your organization. Should it be a:

- Database system,
- Full-text system, or a
- Metadata system?