Annotation for Document Indexing

3S International LLC.

March 1, 2016

Annotation and redaction capabilities are advanced features from document/image viewers. A group of users who share documents among themselves can collaborate via annotations on the contents of the documents. Redactions provide content level securities that enable the protection of sensitive information in portions of a document from users who have access to the entire document. Redaction is a special form of annotation. Many document viewers and image viewers support annotation/redaction capabilities. In ECM repositories, annotation/redaction contents are normally stored separately from the content of the documents that they are associated with.

Such separation has several advantages. First, it allows the documents in the system being annotated or redacted without the content of the documents being modified. Many users can annotate/redact a document simultaneously without worrying about the loss of annotation contents due to the potential concurrency issue.

Second, it allows users annotate/redact documents of different content formats with the same set of annotation/redaction objects. There are many file formats, PDF, MS Word, Text, TIFF, PNG, DWG just to name a few. Without a document viewer, an ECM application must rely on native applications to display the documents in the repository. Some of the applications come with annotation/redaction capabilities. For example, Adobe Acrobat allows users annotate PDF documents. However, Microsoft Office supports a completely different set of annotations. A document viewer control with the annotation/redaction capabilities not only displays documents of many file formats, but also displays a document embedded in web browsers or mobile apps so that users don't have to switch back and forth among different applications, and yet allows users annotate/redact various documents with a predefined set of annotation objects. This is regardless of the native annotation data format that the native applications may have. For example, 3Si hViewer supports the display of documents of many file formats inline and embedded in HTML5 browsers. And hViewer supports the annotate/redact documents of many file formats including PDF and Microsoft Office.

Thirdly, it allows annotation contents sitting on the side of structured-data while all document contents are treated as unstructured-data in the ECM repositories. Searching the structured-data is far more easy and efficient than searching the unstructured-data. With annotation/redaction contents stored separately from the document content, annotations/redactions are qualified as document indexes that are capable of pointing to specific areas of document contents. Imagine having a search hit that points to an annotation

object on page Z of a document, and by clicking the search hit user is taken to a document viewer displaying the page Z of the document where the hit annotation object resides and possibly the annotation object focused and selected in the viewer. This is a huge productivity improvement given that many files in the repository may have multiple pages. An extreme case we had run into in the field was a single PDF file with over 22,000 pages. Providing indexing into a specific page of a multi-page document saves users time and effort from navigating through the pages to look for a search hit.

This is all great. However the question is whether your annotation data is searchable in the repository? If you create a text annotation from a document viewer, and save the annotation object in an ECM repository, is that text searchable? Or if you create an arrow annotation from a document viewer and give a tooltip to the arrow object, is that tooltip text searchable? As the corporate data amounts to the level of BIG DATA, it becomes urgent to answer these questions than ever before. Without being searchable, annotations/redactions are not qualified for document indexing.

Unfortunately, the answer to this question is negative for many commercial ECM products. This is due to the annotation data transparency issue described in another article titled “ECM Data Integrity & Transparency”. To avoid the data transparency issue and other well-known issues associated with the annotation/redaction data in the ECM repositories, some organizations decide to go backwards by restricting document viewers from generating annotation/redaction contents in the repository. This practice is commonly seen among popular EFSS (Enterprise File Synchronization and Sharing) websites such as Dropbox, Box, Syncplicity, Accellion, ShareFile etc. From these cloud storage services, users can view documents from a document viewer. But the document viewer does not support annotation/redaction capabilities. Users simply cannot collaborate on the documents via annotations/redactions.

Disabling annotation/redaction features avoids the issue of the annotation data generated end up not searchable in the repository, and some of other issues. This approach is typically throwing out the baby with the bathwater. Instead of facing the root issue and finding a sound solution, these applications threw away a significant productivity feature to avoid the manifestation of a tough issue. Comparing to the EFSS services that provide document viewer without annotations/redactions, many other EFSS services do not provide a document viewer at all. They rely on native applications to handle the display of the various formats of document contents. We have not yet seen an ECM application that allows users collaborate on documents via annotations/redactions and at the same time makes the annotation/redaction contents searchable outside the document viewer.

Would it be nice if an ECM system allows collaborations via annotations from a document viewer, at the same time enables the search component find hits from annotation data?