pdfSweep

an iText 7 add-on

Keep Your Information Secure

pdfSweep is an iText add-on that removes (redacts) sensitive information from a PDF document. In a secure two-step process, pdfSweep deletes text and images at user-defined coordinates, or as defined by a regular expression. After having parsed the rendering information in the original PDF document, a new PDF document is created without the redacted content. pdfSweep allows you to keep your information secure by creating a new document version that removes sensitive information to ensure complete confidentiality.

Real Redaction

Today's PDF documents represent a significant challenge. In addition to the content, a PDF document contains instructions for rendering the document in a viewer. Adding an instruction to draw a black rectangle does not erase text-rendering instructions underneath the rectangle. This means that 'covering up' information is no longer sufficient, and you will also need to remove or replace the content and relate instructions. Otherwise text extraction would yield the 'redacted' but still available data. pdfSweep provides real redaction.

Key Advantages

- **Remove content instead of just covering it up**
- **Sort content by matching regular expressions**
- **Seamlessly integrate data redaction into your existing workflow**
- **Redact both text and images for complete confidentiality**

Learn more at itextpdf.com/itext7/pdfSweep

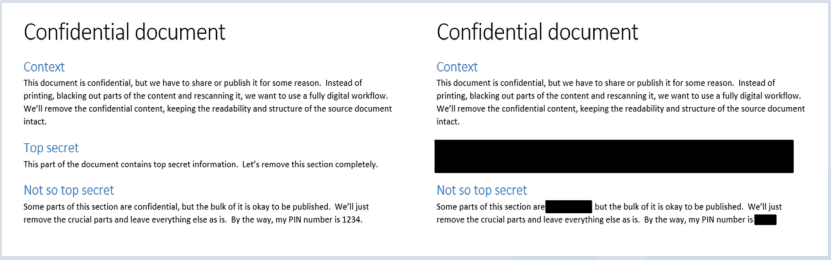
How Does It Work?

The basic **pdfSweep** workflow has just two easy steps. Select those parts of the document that must be redacted: either by specifying the coordinates, or by inputting a regular expression that fits your needs. Then pass the locations to pdfSweep with the pattern of your choice. It is possible to enrich the content by defining a custom color for each snippet of text to be redacted, giving viewers an immediate view of what was redacted, for example red could be a person and green a location.

AUTO SWEEP

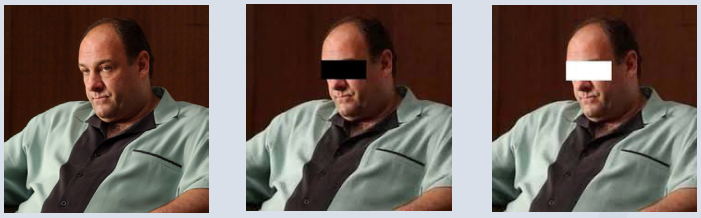
In the **pdfAutoSweep** workflow, the end user can specify a regular expression, and optionally a color. The document is processed a first time, all instructions in the PDF document that relate to text rendering are processed. All characters along with their bounding rectangles in the document are sorted. Next, these intermediate data structures are sorted so that all characters are now in logical (reading) order. The regular expression(s) provided by the user are matched. The information about where the match took place, and the bounding rectangles of the characters involved provide pdfSweep with the rectangles that need to be redacted.

Example



Perform Redaction on Images

Redaction can also be performed on images. The redacted areas are covered with a colored rectangle, and the underlying image is changed (to prevent image extraction from yielding the original image.)



Tony Soprano as displayed in the original PDF document; as displayed after the eyes have been redacted to provide anonymity; extracted image of Tony Soprano (notice that the original content is gone).

Easy and Strong Performance

Most of the regular expressions already configured to work with pdfAutoSweep are inspired by the O'Reilly reference work "Regular Expressions Cookbook." We have already provided a substantial list of common regular expressions to do some of the heavy lifting for you, such as social security numbers, phone numbers, dates, and more. This list will be extended in the future, including expressions related to other national/regional data elements. We invite our community of users to share any expressions they have formulated, in order to enhance the value of the pdfSweep tool. pdfSweep offers high performance, performing linearly in relation to the size of the input document.