

Customer Data Scrambler

InfoPrint Solutions Company

October 8, 2009

Contents

1	Introduction	1
2	Installation	1
2.1	For both AFP and PDF	1
2.2	For PDF only	1
3	AFP	1
3.1	Simple Example	1
3.1.1	Technical Details	2
3.2	Advanced AFP Options	2
4	PDF	2
4.1	Scrambling	2
4.2	PDF Options	3

1 Introduction

Customer Data Scrambler is a tool for scrambling customer-sensitive data inside AFP and PDF files. It is designed for use by the customer to ensure that no sensitive data is inside the print file before sending to InfoPrint Solutions. The customer is responsible for checking the output from this program and ensuring that the data is sufficiently scrambled before sending the file to InfoPrint Solutions.

2 Installation

2.1 For both AFP and PDF

To install the Customer Data Scrambler tool:

- Download the `cds.jar` file from the InfoPrint web site, and place it in a new folder on your system.

That is the only download needed for scrambling the text within an AFP (MO:DCA) file.

2.2 For PDF only

If you wish to scramble text inside a PDF file, additionally download the PDFBox library from the Customer Data Solutions web site; additionally, download and read the Apache license agreement for the PDFBox software. Install the `pdfbox-0.8.0-incubating.jar` and the `commons-logging-1.1.1.jar` files in the same folder as the `cds.jar` file.

3 AFP

3.1 Simple Example

To run the tool to scramble the contents of an AFP file:

- Double-click on the `cds.jar` icon,
- Press the “Choose” button for “Input file” and select the AFP file to be processed,
- Press the “Choose” button for “Output file” and choose an output file,
- Press the “Scramble” button to perform the scrambling. When the process is done, and dialog box will appear with a summary of how much text was processed. To inspire confidence that the tool actually did scramble text, this dialog box has a “Results...” button which shows a list of the first 500 strings of text that were found and what the were scrambled into.

3.1.1 Technical Details

By default, the program will find all normal (i.e. PTOCA and GOCA) text within the AFP document¹. All capital letters are replaced by a random capital letters, all lower-case letters are replaced by random lower-case letters and all digits are replaced by random digits.

Note: In order to properly do its job, all resources must be inline within the print job so the program can examine them all. There are several programs that will take resources in a library and merge them with an AFP file to create a standalone AFP file.

3.2 Advanced AFP Options

The behavior of the program can be changed by pressing the “Advanced options” button, which brings up a new dialog box. This dialog box has two parts. The upper part controls how letters and digits are scrambled. The lower part indicates what to do with TLE’s (Tagged Logical Element structured fields, used for indexing the document), IOCA images, IM images, GOCA graphics, BCOCA bar codes (since there may be logos, pie charts and customer information encoded in these things as well as in text).

For the text options, one can independently choose how capital letters, lower-case letters and digits are scrambled. For each of these, there are four options for changing the text:

- **No change:** The character is not scrambled at all
- **Fixed from:** The character is replaced by a character from the given string, in order. If the string is “xyz” and the text to be scrambled is “abcdefghij” then the result string will be “xyzxyzxyzx.”
- **Sampled from:** The character is replaced by a random character from the given string. If the string is “xyz” and the text to be scrambled is “abcdefghij” then the result string might be “xyzzxyzxyx.”
- **Random from:** The character is replaced by a random character from the given fixed set of characters.

Note: One of the goals of the program is to minimally process the AFP file; if a particular AFP demonstrates a printing problem, the default options will preserve the structure of the file².

4 PDF

4.1 Scrambling

Scrambling a PDF file works just the same as for an AFP file:

¹For those who are technically inclined and who know MO:DCA, the tool finds all IPDS Write Text (WT) and Write Graphics (WG) commands and all MO:DCA Presentation Text (PTX) and Graphics Data (GAD) commands. The only PTOCA bytes that are modified are those outside PTOCA control sequences, or within Transparent Data (TRN) and Repeat String (RPS) control sequences. The only GOCA bytes that are modified are those inside Character String (GCHST) structured fields.

- Double-click on the `cds.jar` icon,
- Press the “Choose” button for “Input file” and select the PDF file to be processed,
- Press the “Choose” button for “Output file” and choose an output file,
- Press the “Scramble” button to perform the scrambling. When the process is done, a dialog box will appear with a summary of how much text was processed.

4.2 PDF Options

The range of options when scrambling text within a PDF document is narrower because of two constraints:

- In keeping with the philosophy of not changing the structure of the print file, the program stores the scrambled text in the exact location that the unscrambled text came from. However, since PDF compresses text, it’s not possible to use random text, since random text doesn’t compress well, and will not fit in the space allocated.
- PDF generally includes only those glyphs from a font that actually appear in the document. This also prevents using random text. The text that is scrambled must come from the glyphs present in the text.

For this reason, there are no user-selectable options for changing the behavior. These constraints mean that the program tried the following in order to scramble the text:

- First it tries to scramble text from code points that exist in the text
- If all else fails, the text is not scrambled.

It is up to the customer to examine the output file to ensure that all customer-sensitive data did indeed get scrambled before sending a PDF file to InfoPrint Solutions.