A tool for redacting the sources: *srcredact*[*]

Boris Veytsman

Revision: 1.4 Date: 2015/05/20 23:30:22

[*]This work was commissioned by the US Consumer Financial Protection Bureau, United States Treasury

Contents

1 Introduction

Many documents containing confidential information exist in several versions: one for the general public, one for the limited audience (or even several confidential versions for the different audiences). In some cases the desire to have several versions of a document might be caused not by confidentiality, but by the different needs of audiences: e.g. an "executive" and "research" version of the same white paper.

One can maintain several versions of the document separately, but this quickly becomes cumbersome and error-prone. At some point the versions drift away too much, and making them close again becomes a difficult and expensive task. This is a typical "anti-pattern", well known to programmers.

Therefore the task is to enable the user to have a single source from which various versions of a document can be produced.

There are different ways to achieve this effect. The *output-level* redaction means that we have a specially marked source file, from which several different PDF files can be produced.

The *source-level* redaction means that we have one source file, also specially marked, from which several different *sources* can be produced. This means that different co-authors can get different versions of a source file and work in their own versions of text. This is the approach taken by the present tool.

One may consider the idea of different co-authors to have different versions of text to be rather strange. Why would an author to be denied the access to a part of her own text? However, there are situations where this idea is warranted. Suppose we have a report that has classified and non-classified parts. Suppose also that some non-classified parts are co-authored by experts that do not have the privilige to read (all or some) classified parts of the report. We want to enable their work on their parts without compromising the confidentiality of the whole.

LaTeX actually provides some facilities for this approach with its `\input` and `\include` mechanism. Indeed, one can put in the document:

```
classified text...

\input{unclassified-part}

classified tex
```

However, this mechanism allows only limited control over the sources.

The program *srcredact* is intended to provide a more fine-grained control over the included and excluded parts.

2 Concept of Operation

Let us first discuss a document, that has two versions: classified and unclassified one. There are two modes for creation of such document with the package *srcredact*. In the first mode the master file is a valid TeX file, which provides the full

```
\documentclass{article}

\begin{document}
Common text
%<*!unclassified>
Classified text
%</!unclassified>
\end{document}
```

Figure 1: Master document in the first mode

```
%<*ALL>
\documentclass{article}

\begin{document}
Common text
%</ALL>
%<*classified>
Classified version of the text
%</!classified>
%<*unclassified>
Unclassified version of the text
%</unclassified>
%<*ALL>
\end{document}
%</ALL>
```

Figure 2: Master document in the second mode

version of the document, and the redacted version is the file with certain parts omitted (Figure 1). In the second mode some parts of the redacted version are *not* present in the unredacted version (Figure 2). These two modes determine two work flows. In the first case flow we run *latex* on the document to get the full version, *or srcredact* and then *latex* to get the redacted (unclassified) version (Figure 3). In the second case we run *srcredact* (with different options) to get either classified or unclassified versions of the package (Figure 4).

In a more complex case we can have more than two versions of the document, for example, intended for different audiences. In all cases we can either plan to a "full" version, obtained by *latex*'ing the master document, or to assume that no version is the superset of all other versions, and thus we need *srcredact* to get any version of the document.

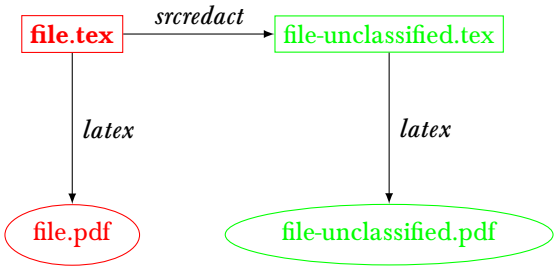

Figure 3: Work flow in the first mode. The red color corresponds to the classified material, the green color to the unclassified one.

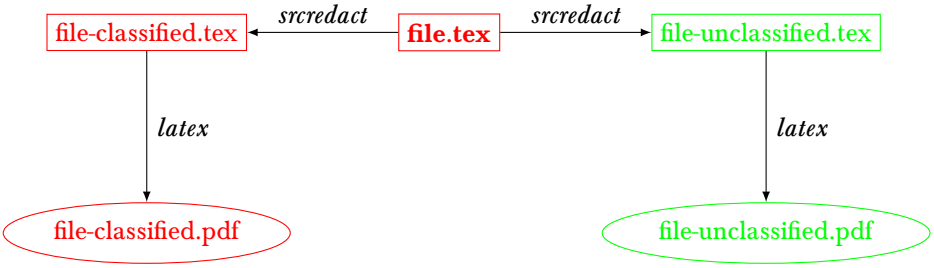

Figure 4: Work flow in the second mode. The red color corresponds to the classified material, the green color to the unclassified one.

3 *srcredact* man page

The following is the manual page of *srcredact* tool.

3.1 NAME

srcredact - a program for redaction of text files

3.2 SYNOPSIS

srcredact [*OPTIONS*] **-e** *audience* [*full_file*]
srcredact [*OPTIONS*] **-u** *audience full_file* [*redacted_file*]
srcredact **-l** [*full_file*]
srcredact **-h**|**-v**

3.3 DESCRIPTION

srcredact is the program to extract "redacted versions" of the master file (option **-e**) or to incorporate the changes in the redacted versions into the master file ("unredact", option **-u**).

The master file consists of *chunks* intended for different audiences. Each audience has a *name*, e.g. `classified`, `unclssified`, `expert` etc. Chunks are started and stopped by *guard lines*. Each guard line has the format (for the default TeX syntax)

```
%<*name1|name2|...>
```

or

```
%</name1|name2|...>
```

In the first cases the text following the guard is *included* for the audiences `name1`, `name2`, ... In the second case it is *excluded* for these audiences.

THere is a special audience `ALL`: a wild card for all audiences. Thus the idiom

```
%</ALL>
%<*classified>
```

means that the chunk is excluded for all audiences but `classified`

Exactly one of the options **-e** (extract) or **-u** (unextract) must be present. In the extract mode the non-option argument is the name of the full file. If it is absent, or is `-`, standard input is used. In the unextract mode the first non-option argument

3·4 OPTIONS

-c ***list of comment patterns***

Use the given pattern for comment lines to search for guards instead of the default `TeX` pattern. The recognized patterns are:

c

```
/*<guard>*/
```

cpp

```
//<guard>
```

fortran

```
C<guard>
```

shell

```
#<guard>
```

TeX

```
%<guard>
```

The pattern names should be separated by commas, and the list may be enclosed in quotes to prevent shell expansion, e.g

```
-c "TeX, c, shell"
```

-d

Debug mode on.

-e ***audience***

Extract the contents for the current audience into the file *file*. The cuurent audience is guessed from the *file* name, if the latter has the structure *base-audience.extension*, e.g. `report-unclassified.tex`. The key **-a** overrides this guess and should be used if the file name does not follow this pattern. The file name - means the standard output.

-h

Print help information and exit.

-l

List all audiences set in the file (one per line) and exit.

-u ***audience***

Take a edited file intended for the *audience* (the second non-option argument) and incorporate the changes in it into the full file (the first non-option argument). If the second argument is missing, standard input is used instead. As usual, - also means standard input. Note that only one of the two file arguments in this case can be standard input.

-v

Print version information and exit.

-w ***on*|*off*|*1*|*0*|*true*|*false***

If `on`, `1` or `true` (the deafult), implicitly wrap the full document into the guards

```
%<*ALL>
...
%</ALL>
```

3.5 RETURN VALUE

The program returns 0 if successful, 1 if conflicts were found in the `unextract` mode and 2 in case of problems.

3.6 CONFLICTS IN UNEXTRACT MODE

Like the standard *diff3*(1) tool, the program may find conflicts between the full version and the edited one in the **-u** mode. Then the resulting file brackets the conflicts in the usual manner, e.g.

```
<<<<<<< /tmp/BrjXo0hMOB/full
%</nobonds>
Forty-five tons best old dry government bonds, suitable for furnace, gold
7 per cents., 1864, preferred.
%<*nobonds>
||||||| /tmp/BrjXo0hMOB/extracted
Forty-five tons best old dry government bonds, suitable for furnace, gold
7 per cents., 1864, preferred.
=======
>>>>>>> /tmp/BrjXo0hMOB/new
```

Here `full` is the full document, `extracted` is the extracted file for the given audience, `new` is the edited file.

3.7 AUTHOR

Boris Veytsman, borisv@lk.net

This work was commissioned by Consumer Financial Protection Bureau, United States Treasury.

3.8 LICENSE AND COPYRIGHT