Technical documentation of rediff tool

David Avsajanishvili

June 18, 2009

Technical documentation of rediff tool
by David Avsajanishvili

Contents

Abstract

`rediff` is a tool for diffing documents of any text-based format by *Sections*, defined using Regular Expressions and marking the *Sections* with custom, format-dependent replacement.

1 Problem definition

1.1 Comparing files

Diff utility is used in Unix-like systems for comparing two text files. It's common usage is tracking changes in files after modification. The utility performs comparison on per-line basis:

- Lines, present in both files are treated as *Unchanged*;
- Lines, present only in current version of the file and absent in the previous are treated as *Added*;
- Lines, present only in previous version of the file and absent in the current are treated as *Removed*;
- If a line is *Removed* and *Added* at the same time, it is treated as *Changed*.

Text files are compared line by line. It means, that an atomic part of the comparison is a *line*, terminated with *newline character sequence* (usually, `\n` or `\r\n`) or end-of-file. Let's call it "Section". So, in plain text file *Section* represents a line of text. It is the simplest way of comparing files and very useful for code sources, configuration files, plain-text documentation, etc.

Although, there are formats for which atomic *Section* can't be represented as just a line. For example:

- In XHTML, XML, DocBook and related formats *Section* is represented as a fragment of text wrapped in *Opening* and *Closing* tags. For example, `<p>Single section</p>`. Also, it is not mandatory that such sections must fit on single line only. Moreover, one section may contain others (example: `<p>Section with <b>bold text</b> inside</p>`).
- In Wiki, AsciiDoc, Markdown and other text-based markup languages *Section* is a fragment of text, defined by the syntax of corresponding language (header, paragraph, etc.).
- In RTF files sections are defined with RTF tags.

More formally, comparing should be possible by any kind of section, specified by format of the file, not only line by line.

1.2 Marking differences in file

When working on text files with presentation different from plain-text (such as HTML or DocBook) it is handy to mark changed *Sections* in it. Marking method is mostly custom and dependent on format used:

- In DocBook differences are marked with `revisionflag` attribute in tags:

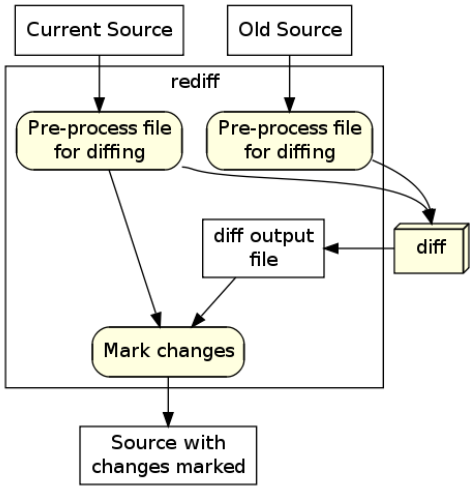
```
<simpara revisionflag="changed">Some changed paragraph.</simpara>
```

- In HTML/XHTML it is useful to mark differences with `class` attribute:

```
<p class="changed">Some changed paragraph.</p>
```

As a result, in *Presentation View* changed sections will be marked, so reader will see which parts of the document were modified since previous revision.

2 Design



The `rediff` uses standard `diff` program for finding differencies. TODO!...