## NAME

**detoxrc -** configuration file for detox(1)

## OVERVIEW

**detox** allows for configuration of its sequences through config files. This document describes how these files work.

## IMPORTANT

When setting up a new set of rules, the *safe* and *wipeup* filters should always be run after a translating filter (or series thereof), such as the *utf_8* or the *uncgi* filters. Otherwise, the risk of introducing difficult characters into the filename is introduced.

## SYNTAX

The format of this configuration file is C-like. It is based loosely off the configuration files used by **named**. Each statement is semicolon terminated, and modifiers on a particular statement are generally contained within braces.

**sequence** "*name*" {*sequence; ...*};

Defines a sequence of filters to run a filename through. *name* specifies how the user will refer to the particular sequence during runtime. Quotes around the sequence name are generally optional, but should be used if the sequence name does not start with a letter.

There is a special sequence, named *default*, which is the default sequence used by **detox**. This can be overridden through the command line option **-s** or the environmental variable DETOX_SEQUENCE.

Sequence names are case sensitive and unique throughout all sequences; that is, if a system-wide file defines *normal_seq* and a user has a sequence with the same name in their *.detoxrc*, the users' *normal_seq* will replace the system-wide version.

**ignore** {**filename** "*filename*"; ...};

Any filename listed here will be ignored during recursion. Note that all files beginning with a period, such as *.git* or *.config* will be ignored by **detox** during recursion.

**# comments**

Any thing after a # on any line is ignored.

### SEQUENCES

All of these statements occur within a **sequence** block.

**iso8859_1**;

**iso8859_1** {**builtin** "*name*";};

**iso8859_1** {**filename** "*/path/to/filename*";};

This transcodes ISO-8859-1 characters between 0xA0 and 0xFF into their UTF-8 equivalents, with a few exceptions. The output is not necessarily safe, and should also be run through the *safe* filter.

If *builtin* is specified, a builtin table with the name specified will be used.

Under normal circumstances, the filename syntax is not needed. **detox** looks in several locations for a file called *iso8859_1.tbl*, which is a set of rules defining how an ISO-8859-1 character should be translated. If **detox** can't find the translation table, it will fall back on the builtin table *iso8859_1*.

You can also download or create your own, and tell **detox** the location of it using the filename syntax shown above.

You can chain together multiple *iso8859_1* filters, as long as the default value of all but the last one it empty. This is explained in detox.tbl(5).

This filter is mutually exclusive with the *utf_8* filter.

**utf_8**;

**utf_8** {**builtin** "*name*";};

**utf_8** {**filename** "*/path/to/filename*";};

This filters Unicode control characters, encoded using UTF-8.

This operates in a manner similar to *iso8859_1*, except it looks for a translation table called *unicode.tbl*.

Similar to the *iso8859_1* filter, an internal table exists, based on the stock translation table, called *unicode*.

**uncgi**;

This translates CGI-escaped strings into their ASCII equivalents. The output of this is not necessarily safe, and should be run through the *safe* filter, at the least.

**safe**;

**safe** {**builtin** "*name*";};

**safe** {**filename** "*/path/to/filename*";};

This could also be called "safe for Unix-like operating systems". It translates characters that are difficult to work with in Unix environments into characters that are not.

Similar to the *iso8859_1* and *utf_8* filters, this can be controlled using a translation table. This filter also has an internal version of the translation table, which can be accessed via the builtin table *safe*.

**wipeup**;

**wipeup** {**remove_trailing**;};

Reduces any series of underscores or dashes to a single character. The dash takes precedence.

If **remove_trailing** is set, then periods are added to the set of characters to work on. The period then takes precedence, followed by the dash.

If a hash character, underscore, or dash are present at the start of the filename, they will be removed.

**max_length** {**length** *value*;};

This trims a filename down to the length specified (or less). It is conscious of extensions and attempts to preserve anything following the last period in a filename.

For instance, given a max length of 12, and a filename of *this_is_my_file.txt*, the filter would output *this_is_.txt*.

**lower**;

This translates uppercase characters into lowercase characters. It only works on ASCII characters.

## BUILTIN TABLES

cp1252

A translation table for transcoding CP-1252 characters to UTF-8, with a few exceptions. This is no longer a common use case, and has been moved to a separate table.

iso8859_1

A translation table for transcoding single-byte characters with the high bit set from ISO-8859-1 to

UTF-8.

safe A replacement table for characters that are hard to work with under Unix and Unix-like OSs.

unicode
A translation table for converting multi-byte control characters encoded in UTF-8 to safe characters.

## EXAMPLES

```
# filter UTF-8 control characters to ASCII (using chained tables), clean up
sequence utf8 {
  utf_8 {
    filename "/usr/local/share/detox/custom.tbl";
  };
  utf_8 {
    builtin "unicode";
  };
  safe {
    builtin "safe";
  };
  wipeup {
    remove_trailing;
  };
  max_length {
    length 128;
  };
};
# decode CGI, transcode CP-1252 to UTF-8, clean up
sequence "cgi-cp1252" {
  uncgi;
  iso8859_1 {
    builtin "cp1252";
  };
  safe {
    builtin "safe";
  };
};
```

## SEE ALSO

detox(1), inline-detox(1), detox.tbl(5), ascii(7), iso_8859-1(7), unicode(7), utf-8(7)

## AUTHORS

detox was written by Doug Harple.