

# Automated Documentation Proofreading

Warren Block  
wblock@freebsd.org  
FreeBSD documentation committer

# Why is Documentation Hard to Write?

# Why is Documentation Hard to Write?

Rules, rules, so many different rules!

# Why is Documentation Hard to Write?

Rules, rules, so many different rules!

Text files

# Why is Documentation Hard to Write?

Rules, rules, so many different rules!

Text files

mdoc ( 7 )

# Why is Documentation Hard to Write?

Rules, rules, so many different rules!

Text files

mdoc ( 7 )

DocBook SGML

# Why is Documentation Hard to Write?

**Existing documentation is inconsistent**

Learning by example is difficult when the examples vary wildly in quality

# Why is Documentation Hard to Write?

## Toolchains are unhelpful

Format and style errors are often unreported

```
<para>The new setting may be viewed as before,  
notice the <literal>s</literal> is now in the  
field designated for the group permission  
settings:<para>
```



# Why is Documentation Hard to Write?

## Toolchains are unhelpful

Format and style errors are often unreported

```
<para>The new setting may be viewed as before,  
notice the <literal>s</literal> is now in the  
field designated for the group permission  
settings:</para>
```



# Why Worry? If It Builds, Ship It!

## Quality

Consistency encourages quality

# Why Worry? If It Builds, Ship It!

## Maintenance

Clean, consistent documents are easier to understand, maintain, and modify

# Why Worry? If It Builds, Ship It!

## Conversion To Other Formats

DocBook XML

mandoc

future formats

# Why Worry? If It Builds, Ship It!

## Entropy

Problems accumulate

# Why Worry? If It Builds, Ship It!

## Entropy

*The FreeBSD Porter's Handbook*

16,000 lines of DocBook SGML

# Why Worry? If It Builds, Ship It!

## Entropy

*The FreeBSD Porter's Handbook*

16,000 lines of DocBook SGML

To fix whitespace required an 8,000-line commit.

# Why Worry? If It Builds, Ship It!

## Entropy

*The FreeBSD Porter's Handbook*

16,000 lines of DocBook SGML

To fix whitespace required an 8,000-line commit.

*Followed by another 4,000-line commit.*



# What Can Be Done?

**Make things easier for writers!**

*Especially* for people who rarely work on documentation at all.

# What Can Be Done?

**Make things easier for writers!**

*Especially* for people who rarely work on documentation at all.

**Encourage  
programmers to  
document their work**

# What Can Be Done?

**Make things easier for writers!**

*Especially* for people who rarely work on documentation at all.

Encourage  
programmers to  
document their work

Encourage users to  
improve the quality  
of documentation

# What Can Be Done?

**Make things easier for writers!**

*Especially* for people who rarely work on documentation at all.

Encourage programmers to document their work

Encourage users to improve the quality of documentation

Encourage writers to expand and clarify documentation

# What Can Be Done?

## Automated Proofreading

# What Can Be Done?

## Automated Proofreading

Remember things

Help those who suffer from CRS syndrome

# What Can Be Done?

## Automated Proofreading

Remember things

Find errors

Subtle errors, but also errors of inexperience

# What Can Be Done?

## Automated Proofreading

Remember things

Find errors

Help comply with standards

Indirectly educate the user on standards



# What Can Be Done?

## Automated Proofreading

Remember things

Find errors

Help comply with standards

Keep mistakes out of the tree

# What Can Be Done?

## Automated Proofreading

Remember things

Find errors

Help comply with standards

Keep mistakes out of the tree

Let the writer concentrate on the message!

# What Tests Can Be Automated?

Tests For All Files

# What Tests Can Be Automated?

## Tests For All Files

Spelling

Use misspellings from FreeBSD text files, man pages, and DocBook source

# What Tests Can Be Automated?

## Tests For All Files

Spelling

Repeated words

Detect repeated words in a line or from one line to the next

# What Tests Can Be Automated?

## Tests For All Files

Spelling

Repeated words

Bad phrases

“The to”, “to for”, again from actual FreeBSD files

# What Tests Can Be Automated?

## Tests For All Files

### Writing Style

*you* and *your*  
*should*  
*obviously* and *needless to say*  
*simply* and *basically*  
starting too many sentences with *the*  
*e.g.* and *i.e.*  
No examples!

Great potential to improve readability and clarity

# What Tests Can Be Automated?

`mdoc(7)` Tests



# What Tests Can Be Automated?

## `mdoc(7)` Tests

Sentences begin on a new line

Not enforced by the toolchain

# What Tests Can Be Automated?

## mdoc ( 7 ) Tests

Sentences begin on a new line

Document date updated on non-trivial changes

Easy to forget

# What Tests Can Be Automated?

## `mdoc(7)` Tests

Sentences begin on a new line

Document date updated on non-trivial changes

Structure: the eight minimum macros

# What Tests Can Be Automated?

## `mdoc(7)` Tests

Sentences begin on a new line

Document date updated on non-trivial changes

Structure: the eight minimum macros

```
.Dd  
.Dt  
.Os  
.Sh NAME
```

```
.Nm  
.Nd  
.Sh SYNOPSIS  
.Sh DESCRIPTION
```

From the manual page template in `mdoc(7)`

# What Tests Can Be Automated?

## DocBook SGML Tests

Rules are described in the *FreeBSD Documentation Project Primer*

# What Tests Can Be Automated?

## DocBook SGML Tests

Whitespace

Tabs versus spaces at the beginning of lines

# What Tests Can Be Automated?

## DocBook SGML Tests

Whitespace

Indentation

Indent level, matching open/close tags, lines wrap at 70 columns

# What Tests Can Be Automated?

## DocBook SGML Tests

Whitespace

Indentation

Tag usage style

Tags like `<programlisting>` need special handling



# What Tests Can Be Automated?

## DocBook SGML Tests

Whitespace

Indentation

Tag usage style

Title capitalization

# Igor, The Lab Assistant

# **igor, The Lab Assistant**

**Must be easy and quick to use**

# **igor, The Lab Assistant**

**Must be easy and quick to use**

Auto-detect type of input file

# **igor, The Lab Assistant**

**Must be easy and quick to use**

Auto-detect type of input file

Handle multiple files and compressed files

# igor, The Lab Assistant

**Must be easy and quick to use**

Auto-detect type of input file

Handle multiple files and compressed files

Test for conformance with the *FDP Primer*

# igor, The Lab Assistant

**Must be easy and quick to use**

Auto-detect type of input file

Handle multiple files and compressed files

Test for conformance with the *FDP Primer*

Be able to run one test or all

# igor, The Lab Assistant

**Must be easy and quick to use**

Auto-detect type of input file

Handle multiple files and compressed files

Test for conformance with the *FDP Primer*

Be able to run one test or all

Avoid false positives



# **igor, The Lab Assistant**

## **Implementation**

# **igor, The Lab Assistant**

## **Implementation**

Written in Perl, but whatever

# igor, The Lab Assistant

## Implementation

Written in Perl, but whatever  
It's regexes all the way down

# **igor, The Lab Assistant**

**What does it look like?**

# igor, The Lab Assistant

## Checking man pages

```
% igor -D man1/kgdb.1.gz man1/link.1.gz man1/lockf.1.gz man8/newsyslog.8.gz man8/boot.8.gz man8/atmconfig.8.gz man8/boot0cfg.8.gz man8/flowctl.8.gz | less -RS
man1/kgdb.1.gz:
kgdb.1.gz:102:bad phrase:[to for] a remote debugging session.
man1/link.1.gz:
link.1.gz:232:spelling:[extention] and should not be used in portable scripts.
man1/lockf.1.gz:
lockf.1.gz:142:trailing whitespace:The[ ]
man8/newsyslog.8.gz:
newsyslog.8.gz:170:repeated:will not [be be] automatically removed (unless the new format is very
man8/boot.8.gz:
boot.8.gz:45:Sh DESCRIPTION used here:but .Sh SYNOPSIS has not been defined
man8/atmconfig.8.gz:
atmconfig.8.gz:4:tab after space:.\ "[ ]All rights reserved.
man8/boot0cfg.8.gz:
boot0cfg.8.gz:109:sentence not on new line:0x1b8 in the MBR. [This information is sometimes used by NT, XP
boot0cfg.8.gz:110:sentence not on new line:to identify the disk drive. [The option is only compatible with
boot0cfg.8.gz:168:sentence not on new line:manpage. [Specifically, do a]
boot0cfg.8.gz:217:sentence not on new line:content. [Be careful.]
man8/flowctl.8.gz:
flowctl.8.gz:35:Sh SYNOPSIS used here:but .Nd has not been defined
flowctl.8.gz:39:Sh DESCRIPTION used here:but .Nd has not been defined
```

**-D to skip comparing .Dd with today's date**

# igor, The Lab Assistant

## Clarifying the output

```
% igor -R -D man1/kgdb.1.gz man1/link.1.gz man1/lockf.1.gz man8/newsyslog.8.gz man8/boot.8.gz man8/atmconf
ig.8.gz man8/boot0cfg.8.gz man8/flowctl.8.gz | less -RS
man1/kgdb.1.gz:
kgdb.1.gz:102:bad phrase:to for a remote debugging session.
man1/link.1.gz:
link.1.gz:232:spelling:extention and should not be used in portable scripts.
man1/lockf.1.gz:
lockf.1.gz:142:trailing whitespace:The
man8/newsyslog.8.gz:
newsyslog.8.gz:170:repeated:will not be be automatically removed (unless the new format is very
man8/boot.8.gz:
boot.8.gz:45:Sh DESCRIPTION used here:but .Sh SYNOPSIS has not been defined
man8/atmconfig.8.gz:
atmconfig.8.gz:4:tab after space:.\ "All rights reserved.
man8/boot0cfg.8.gz:
boot0cfg.8.gz:109:sentence not on new line:0x1b8 in the MBR. This information is sometimes used by NT, XP
boot0cfg.8.gz:110:sentence not on new line:to identify the disk drive. The option is only compatible with
boot0cfg.8.gz:168:sentence not on new line:manpage. Specifically, do a
boot0cfg.8.gz:217:sentence not on new line:content. Be careful.
man8/flowctl.8.gz:
flowctl.8.gz:35:Sh SYNOPSIS used here:but .Nd has not been defined
flowctl.8.gz:39:Sh DESCRIPTION used here:but .Nd has not been defined
```

**-R to produce ANSI color sequences**

# igor, The Lab Assistant

## Checking writing style

```
% igor -R -y chapter.sgml
chapter.sgml style check:
"you" used 512 times
"your" used 156 times
  "You" and "your" are informal and subjective.
  Try to be formal and objective: "the file" rather than "your file".
"should" used 37 times
  Use "should" sparingly, it is feeble.
  Try to be imperative: "do this" rather than "you should do this".
"obviously" used 1 time
  If it is really obvious, it does not need to be pointed out.
"simply" used 13 times
"basically" used 1 time.
  Can be read as patronizing.
"e.g." used 6 times
  "E.g." (Latin "exempli gratia") means "for example" and is mostly
  used in academic and scientific writing. Consider replacing with the
  more common English words. Both forms are usually followed by a
  comma for a verbal pause: "e.g., a b c" or "for example, a b c"
"i.e." used 1 time
  "I.e." (Latin "id est") means "that is" and is mostly used in academic
  and scientific writing. Consider replacing with the more common
```

-y for style tests

# igor, The Lab Assistant

## Checking DocBook whitespace

```
% igor -RZ advanced-networking/chapter.sgml | less -RS
chapter.sgml:24:wrap long line:      <para>How to set up &ieee; 802.11 and &bluetooth; devices.</para>
chapter.sgml:32:wrap long line:      <para>How to set up network booting on a diskless machine.</para>
chapter.sgml:36:wrap long line:      <para>How to set up network PXE booting with an NFS root filesystem
chapter.sgml:65:wrap long line:      <para>Understand the basics of the <filename>/etc/rc</filename> scrip
chapter.sgml:73:use tabs instead of spaces: <para>Know how to configure and install a new FreeBSD k
chapter.sgml:74:use tabs instead of spaces:      (<xref linkend="kernelconfig">).</para>
chapter.sgml:78:bad tag indent: <para>Know how to install additional third-party
chapter.sgml:79:use tabs instead of spaces:      software (<xref linkend="ports">).</para>
chapter.sgml:80:bad tag indent: </listitem>
chapter.sgml:88:use tabs instead of spaces: <author>
chapter.sgml:89:use tabs instead of spaces:      <firstname>Coranth</firstname>
chapter.sgml:90:tab after space: <surname>Gryphon</surname>
chapter.sgml:91:bad tag indent: <contrib>Contributed by </contrib>
chapter.sgml:92:use tabs instead of spaces: </author>
chapter.sgml:115:stragglng </para>: </para>
chapter.sgml:120:wrap long line:      <para>To illustrate different aspects of routing, we will use the
chapter.sgml:141:wrap long line:      section</link>) and the <hostid>localhost</hostid> route.</para>
chapter.sgml:153:use tabs instead of spaces: <primary>Ethernet</primary>
chapter.sgml:154:use tabs instead of spaces: <secondary>MAC address</secondary>
chapter.sgml:172:wrap long line:      <para>FreeBSD will also add subnet routes for the local subnet (<hos
chapter.sgml:173:wrap long line:      role="ipaddr">10.20.30.255</hostid> is the broadcast address for
```

-z for whitespace tests only



# igor, The Lab Assistant

## Checking DocBook content

```
% igor -Rz disks/chapter.sgml | less -RS
chapter.sgml:810:no comma after "e.g.":   e.g. by enabling <literal>vfs.usermount</literal> as
chapter.sgml:986:capitalization:         <title>mkisofs</title>
chapter.sgml:1067:capitalization:        <title>burncd</title>
chapter.sgml:1088:capitalization:        <title>cdrecord</title>
chapter.sgml:1931:spelling:              <para>A floppy disk needs to be low-level formatted before it
chapter.sgml:1980:spelling:              <para>Now the floppy is ready to be high-level formatted. This
chapter.sgml:2345:no comma after "e.g.": <para>Complete machine destruction (e.g. fire), including
chapter.sgml:2375:no comma after "e.g.": <para>Copies of whole filesystems and/or disks (e.g. perio
chapter.sgml:2492:capitalization:        <title>Using <command>dump</command> over <application>ssh</applic
chapter.sgml:2674:no comma after "e.g.": (e.g. <command>bsdlab
chapter.sgml:2743:no comma after "e.g.": <para>Try to <command>mount</command> (e.g. <command>mount
chapter.sgml:2750:no comma after "e.g.": recover the data for this file system (e.g. <command>res
chapter.sgml:2751:no comma after "e.g.": /dev/sa0</command>). Unmount the file system (e.g. <c
chapter.sgml:3018:no comma after "e.g.": <para>One can find snapshot files on a file system (e.g. <fi
chapter.sgml:3433:capitalization:        <title>Create a Directory to Hold gbde Lock Files</title>
chapter.sgml:3448:capitalization:        <title>Initialize the gbde Partition</title>
chapter.sgml:3483:open <para> without closing: <para>The <command>gbde init</command> command creates a
chapter.sgml:3583:capitalization:        <title>Attach the gbde Partition to the Kernel</title>
chapter.sgml:3639:capitalization:        <title>Cryptographic Protections Employed by gbde</title>
chapter.sgml:3705:no comma after "e.g.": <para>Allows the use of two independent keys (e.g. a
chapter.sgml:3798:capitalization:        <title>Attaching the Provider with the generated Key</title>
```

-z for non-whitespace tests only

# Where Is It?

`/usr/ports/textproc/igor`

`http://www.wonkity.com/~wblock/igor/`

# Lessons Learned

# Lessons Learned

Optimize regexes, short-circuit when possible

# Lessons Learned

Optimize regexes, short-circuit when possible  
DocBook SGML indentation is non-trivial

# Lessons Learned

Optimize regexes, short-circuit when possible

DocBook SGML indentation is non-trivial

Syntax highlighting is good for whitespace

# Lessons Learned

Optimize regexes, short-circuit when possible

DocBook SGML indentation is non-trivial

Syntax highlighting is good for whitespace

Advertising

# The Future



# The Future

Rewrite

Better language, or style, or speed

# The Future

Rewrite

Better DocBook indentation testing

Something smarter, maybe even a full parser

# The Future

Rewrite

Better DocBook indentation testing

Advanced language analysis

Analyze content rather than just words

# The Future

Rewrite

Better DocBook indentation testing

Advanced language analysis

Other languages

Possibly easier than it sounds

# The End

Special thanks to  
Glen Barber and Benedict Reuschling



<http://www.wonkity.com/~wblock/igor>

Thank you!