Scsh Reference Manual

For scsh release 0.5.2

Olin Shivers and Brian D. Carlstrom

Acknowledgements

Who should I thank? My so-called "colleagues," who laugh at me behind my back, all the while becoming famous on *my* work? My worthless graduate students, whose computer skills appear to be limited to downloading bitmaps off of netnews? My parents, who are still waiting for me to quit "fooling around with computers," go to med school, and become a radiologist? My department chairman, a manager who gives one new insight into and sympathy for disgruntled postal workers?

My God, no one could blame me—no one!—if I went off the edge and just lost it completely one day. I couldn't get through the day as it is without the Prozac and Jack Daniels I keep on the shelf, behind my Tops-20 JSYS manuals. I start getting the shakes real bad around 10am, right before my advisor meetings. A 10 oz. Jack 'n Zac helps me get through the meetings without one of my students winding up with his severed head in a bowling-ball bag. They look at me funny; they think I twitch a lot. I'm not twitching. I'm controlling my impulse to snag my 9mm Sig-Sauer out from my day-pack and make a few strong points about the quality of undergraduate education in Amerika.

If I thought anyone cared, if I thought anyone would even be reading this, I'd probably make an effort to keep up appearances until the last possible moment. But no one does, and no one will. So I can pretty much say exactly what I think.

Oh, yes, the *acknowledgements.* I think not. I did it. I did it all, by myself.

Olin Shivers
Cambridge
September 4, 1994

Contents

1 Introduction **1**

1.1 Obtaining scsh . . . 1

1.2 Building scsh . . . 2

1.3 Caveats . . . 2

1.4 Naming conventions . . . 3

1.5 Lexical issues . . . 4

1.5.1 Extended symbol syntax . . . 4

1.5.2 Extended string syntax . . . 5

1.5.3 Block comments and executable interpreter-triggers . . . 5

1.5.4 Here-strings . . . 5

1.5.5 Dot . . . 7

1.6 Record types and the `define-record` form . . . 7

1.7 A word about Unix standards . . . 9

2 Process notation **10**

2.1 Extended process forms and `i/o` redirections . . . 10

2.1.1 Port and file descriptor sync . . . 11

2.2 Process forms . . . 12

2.3 Using extended process forms in Scheme . . . 13

2.3.1 Procedures and special forms . . . 14

2.3.2 Interfacing process output to Scheme . . . 14

2.4 More complex process operations . . . 16

2.4.1 Pids and ports together . . . 17

2.4.2 Multiple stream capture . . . 17

2.5 Conditional process sequencing forms . . . 19

2.6 Process filters 20

3 System Calls **21**

3.1 Errors 21
3.1.1 Interactive mode and error handling 23
3.2 I/O 24
3.2.1 Standard R4RS I/O procedures 24
3.2.2 Port manipulation and standard ports 24
3.2.3 String ports 26
3.2.4 Revealed ports and file descriptors 26
3.2.5 Port-mapping machinery 29
3.2.6 Unix I/O 31
3.2.7 Buffered I/O 36
3.2.8 File locking 37
3.3 File system 40
3.4 Processes 52
3.4.1 Process objects and process reaping 55
3.4.2 Process waiting 58
3.4.3 Analysing process status codes 59
3.5 Process state 60
3.6 User and group database access 61
3.7 Accessing command-line arguments 62
3.8 System parameters 64
3.9 Signal system 64
3.10 Time 68
3.10.1 Terminology 68
3.10.2 Basic data types 69
3.10.3 Time zones 70
3.10.4 Procedures 70
3.11 Environment variables 74
3.11.1 Path lists and colon lists 76
3.11.2 `$USER`, `$HOME`, and `$PATH` 77
3.12 Terminal device control 77
3.12.1 Portability across OS variants 78
3.12.2 Miscellaneous procedures 78

3.12.3 The tty-info record type 78
3.12.4 Using tty-info records 80
3.12.5 Other terminal-device procedures 81
3.12.6 Control terminals, sessions, and terminal process groups 82
3.12.7 Pseudo-terminals 83

4 Networking 90
4.1 High-level interface 90
4.2 Sockets 91
4.3 Socket addresses 92
4.4 Socket primitives 93
4.5 Performing input and output on sockets 95
4.6 Socket options 96
4.7 Database-information entries 97

5 Strings and characters 99
5.1 String manipulation 99
5.1.1 Manipulating file-names 99
5.1.2 Other string manipulation facilities 105
5.2 ASCII encoding 106
5.3 Character sets 106
5.3.1 Creating character sets 107
5.3.2 Querying character sets 108
5.3.3 Character-set algebra 108
5.3.4 Standard character sets 109
5.3.5 Linear-update character-set operations 110

6 Pattern-matching strings with regular expressions 112
6.1 Summary SRE syntax 113
6.2 Examples 117
6.3 A short tutorial 117
6.4 Choices 119
6.4.1 Embedding regexps within Scheme programs 128
6.5 Regexp functions 129
6.5.1 Obsolete, deprecated procedures 129
6.5.2 Standard procedures and syntax 129

6.6 The regexp ADT . . . 136
6.7 Syntax-hacking tools . . . 138

7 Reading delimited strings **140**

8 Awk, record I/O, and field parsing **143**
8.1 Record I/O and field parsing . . . 143
8.1.1 Reading records . . . 144
8.1.2 Parsing fields . . . 144
8.1.3 Field readers . . . 148
8.1.4 Forward-progress guarantees and empty-string matches 148
8.1.5 Reader limitations . . . 150
8.2 Awk . . . 150
8.2.1 Examples . . . 153
8.3 Backwards compatibility . . . 155

9 Miscellaneous routines **156**
9.1 Integer bitwise ops . . . 156
9.2 List procedures . . . 156
9.3 Top level . . . 156

10 Running scsh **158**
10.1 Scsh command-line switches . . . 159
10.1.1 Scripts and programs . . . 159
10.1.2 Inserting interpreter triggers into scsh programs . . . 159
10.1.3 Module system . . . 160
10.1.4 Switches . . . 161
10.1.5 The meta argument . . . 165
10.1.6 Examples . . . 167
10.1.7 Process exit values . . . 169
10.2 The scsh virtual machine . . . 170
10.2.1 VM arguments . . . 170
10.2.2 Inserting interpreter triggers into heap images . . . 171
10.2.3 Inserting a double-level trigger into Scheme programs . 172
10.3 Compiling scsh programs . . . 172
10.4 Statically linking heap images . . . 173

10.5 Standard file locations 175

11 Changes from previous releases **177**

11.1 Changes from the previous release 177

11.2 Changes from the penultimate release 178

11.2.1 New features 178

11.2.2 Incompatible improvements 179

11.2.3 Backwards-compatible improvements 182

12 Todo **183**

Chapter 1

Introduction

This is the reference manual for scsh, a Unix shell that is embedded within Scheme. Scsh is a Scheme system designed for writing useful standalone Unix programs and shell scripts—it spans a wide range of application, from "script" applications usually handled with perl or sh, to more standard systems applications usually written in C.

Scsh comes built on top of Scheme 48, and has two components: a process notation for running programs and setting up pipelines and redirections, and a complete syscall library for low-level access to the operating system. This manual gives a complete description of scsh. A general discussion of the design principles behind scsh can be found in a companion paper, "A Scheme Shell."

1.1 Obtaining scsh

Scsh is distributed via net publication. We place new releases at well-known network sites, and allow them to propagate from there. We currently release scsh to the following Internet sites:

```
ftp://ftp-swiss.ai.mit.edu/pub/su/
http://www-swiss.ai.mit.edu/scsh/scsh.html
http://www.cs.indiana.edu/scheme-repository/
```

These sites are the MIT Project Mac ftp server, the Scheme Shell home page, and the Indiana Scheme Repository home page, respectively. Each should have a compressed tar file of the entire scsh release, which includes all the source code and the manual, and a separate file containing just this manual in Postscript form, for those who simply wish to read about the system.

However, nothing is certain for long on the Net. Probably the best way to get a copy of scsh is to use a network resource-discovery tool, such as archie, to find ftp servers storing scsh tar files. Take the set of sites storing the most recent release of scsh, choose one close to your site, and download the tar file.

1.2 Building scsh

Scsh currently runs on a fairly large set of Unix systems, including Linux, NetBSD, SunOS, Solaris, AIX, NeXTSTEP, Irix, and HP-UX. We use the Gnu project's autoconfig tool to generate self-configuring shell scripts that customise the scsh Makefile for different OS variants. This means that if you use one of the common Unix implementations, building scsh should require exactly the following steps:

`gunzip scsh.tar.gz`	*Uncompress the release tar file.*
`untar xfv scsh.tar`	*Unpack the source code.*
`cd scsh-0.5`	*Move to the source directory.*
`./configure`	*Examine host; build Makefile.*
`make`	*Build system.*

When you are done, you should have a virtual machine compiled in file `scshvm`, and a heap image in file `scsh/scsh.image`. Typing

```
make install
```

will install these programs in your installation directory (by default, `/usr/local`), along with a small stub startup binary, `scsh`.

If you don't have the patience to do this, you can start up a Scheme shell immediately after the initial make by simply saying

```
./scshvm -o ./scshvm -i scsh/scsh.image
```

See chapter 10 for full details on installation locations and startup options.

It is not too difficult to port scsh to another Unix platform if your OS is not supported by the current release. See the release notes for more details on how to do this.

1.3 Caveats

It is important to note what scsh is *not*, as well as what it is. Scsh, in the current release, is primarily designed for the writing of shell scripts—programming. It is not a very comfortable system for interactive command use: the current release lacks job control, command-line editing, a terse, convenient command syntax, and it does not read in an initialisation file analogous to `.login` or `.profile`. We hope to address all of these issues in future releases; we even have designs for several of these features; but the system as-released does not currently provide these features.

In the current release, the system has some rough edges. It is quite slow to start up—loading the initial image into the Scheme 48 virtual machine takes about a cpu second. This can be fixed with the static heap linker provided with this release.

This manual is very, very rough. At some point, we hope to polish it up, finish it off, and re-typeset it using markup, so we can generate html, info nodes, and TeX output from the single source without having to deal with Texinfo. But it's all there is, for now.

1.4 Naming conventions

Scsh follows a general naming scheme that consistently employs a set of abbreviations. This is intended to make it easier to remember the names of things. Some of the common ones are:

`fdes` Means "file descriptor," a small integer used in Unix to represent I/O channels.

`...*` A given bit of functionality sometimes comes in two related forms, the first being a *special form* that contains a body of Scheme code to be executed in some context, and the other being a *procedure* that takes a procedural argument (a "thunk") to be called in the same context. The procedure variant is named by taking the name of the special form, and appending an asterisk. For example:

```
;;; Special form:
(with-cwd "/etc"
  (for-each print-file (directory-files))
  (display "All done"))

;;; Procedure:
(with-cwd* "/etc"
  (lambda ()
    (for-each print-file (directory-files))
    (display "All done")))
```

action/*modifier* The infix "/" is pronounced "with," as in `exec/env`—"exec with environment."

`call/...` Procedures that call their argument on some computed value are usually named "`call/...`," *e.g.*, `(call/fdes` *port proc*`)`, which calls *proc* on *port*'s file descriptor, returning whatever *proc* returns. The abbreviated name means "call with file descriptor."

`with-...` Procedures that call their argument, and special forms that execute their bodies in some special dynamic context frequently have names of the form `with-...`. For example, `(with-env` *env* $body_1$ `...)` and `(with-env*` *env thunk*`)`. These forms set the process environment body, execute their body or thunk, and then return after resetting the environment to its original state.

`create-` Procedures that create objects in the file system (files, directories, temp files, fifos, *etc.*), begin with `create-...`.

`delete-` Procedures that delete objects from the file system (files, directories, temp files, fifos, *etc.*), begin with `delete-...`.

record`:`*field* Procedures that access fields of a record are usually written with a colon between the name of the record and the name of the field, as in `user-info:home-dir`.

`%...` A percent sign is used to prefix lower-level scsh primitives that are not commonly used.

`-info` Data structures packaging up information about various OS entities frequently end in ...`-info`. Examples: `user-info`, `file-info`, `group-info`, and `host-info`.

Enumerated constants from some set *s* are usually named $s/const_1$, $s/const_2$, ... For example, the various Unix signal integers have the names `signal/cont`, `signal/kill`, `signal/int`, `signal/hup`, and so forth.

1.5 Lexical issues

Scsh's lexical syntax is just R4RS Scheme, with the following exceptions.

1.5.1 Extended symbol syntax

Scsh's symbol syntax differs from R4RS Scheme in the following ways:

- In scsh, symbol case is preserved by `read` and is significant on symbol comparison. This means

```
(run (less Readme))
```

displays the right file.

- "`-`" and "`+`" are allowed to begin symbols. So the following are legitimate symbols:

```
-O2 -geometry +Wn
```

- "`|`" and "`.`" are symbol constituents. This allows `|` for the pipe symbol, and `..` for the parent-directory symbol. (Of course, "`.`" alone is not a symbol, but a dotted-pair marker.)

- A symbol may begin with a digit. So the following are legitimate symbols:

```
9x15 80x36-3+440
```

1.5.2 Extended string syntax

Scsh strings are allowed to contain the ANSI C escape sequences such as `\n` and `\161`.

1.5.3 Block comments and executable interpreter-triggers

Scsh allows source files to begin with a header of the form

```
#!/usr/local/bin/scsh -s
```

The Unix operating system treats source files beginning with the headers of this form specially; they can be directly executed by the operating system (see chapter 10 for information on how to use this feature). The scsh interpreter ignores this special header by treating `#!` as a comment marker similar to `;`. When the scsh reader encounters `#!`, it skips characters until it finds the closing sequence newline/exclamation-point/sharp-sign/newline.

Although the form of the `#!` read-macro was chosen to support interpreter-triggers for executable Unix scripts, it is a general block-comment sequence and can be used as such anywhere in a scsh program.

1.5.4 Here-strings

The read macro `#<` is used to introduce "here-strings" in programs, similar to the `<<` "here document" redirections provided by sh and csh. There are two kinds of here-string, character-delimited and line-delimited; they are both introduced by the `#<` sequence.

Character-delimited here-strings

A *character-delimited* here-string has the form

```
#<x...stuff...x
```

where *x* is any single character (except `<`, see below), which is used to delimit the string bounds. Some examples:

<table>
<tr><th>Here-string syntax</th><th>Ordinary string syntax</th></tr>
<tr><td><code>#<|Hello, world.|</code></td><td><code>"Hello, world."</code></td></tr>
<tr><td><code>#<!"Ouch," he said.!</code></td><td><code>"\"Ouch,\" he said."</code></td></tr>
</table>

There is no interpretation of characters within the here-string; the characters are all copied verbatim.

Line-delimited here-strings

If the sequence begins "`#<<`" then it introduces a *line-delimited* here-string. These are similar to the "here documents" of sh and csh. Line-delimited here-strings are delimited by the rest of the text line that follows the "`#<<`" sequence. For example:

```
#<<FOO
Hello, there.
This is read by Scheme as a string,
terminated by the first occurrence
of newline-F-O-O-newline or newline-F-O-O-eof.
FOO
```

Thus,

```
#<<foo
Hello, world.
foo
```

is the same thing as

```
"Hello, world."
```

Line-delimited here-strings are useful for writing down long, constant strings—such as long, multi-line `format` strings, or arguments to Unix programs, *e.g.*,

```
;; Free up some disk space for my netnews files.
(run (csh -c #<<EOF
cd /urops
rm -rf *
echo All done.

EOF
))
```

The advantage they have over the double-quote syntax (*e.g.*, `"Hello, world."`) is that there is no need to backslash-quote special characters internal to the string, such as the double-quote or backslash characters.

The detailed syntax of line-delimited here-strings is as follows. The characters "`#<<`" begin the here-string. The characters between the "`#<<`" and the next newline are the *delimiter line*. All characters between the "`#<<`" and the next newline comprise the delimiter line—including any white space. The body of the string begins on the following line, and is terminated by a line of text which exactly matches the delimiter line. This terminating line can be ended by either a newline or end-of-file. Absolutely no interpretation is done on the input string. Control characters, white space, quotes, backslash—everything is copied as-is. The newline immediately preceding the terminating delimiter line is not included in the result string (leave an extra blank line if you need to put a final newline in the here-string—see the example above). If EOF is encountered before reading the end of the here-string, an error is signalled.

1.5.5 Dot

It is unfortunate that the single-dot token, "`.`", is both a fundamental Unix file name and a deep, primitive syntactic token in Scheme—it means the following will not parse correctly in scsh:

```
(run/strings (find . -name *.c -print))
```

You must instead quote the dot:

```
(run/strings (find "." -name *.c -print))
```

When you write shell scripts that manipulate the file system, keep in mind the special status of the dot token.

1.6 Record types and the `define-record` form

Scsh's interfaces occasionally provide data in structured record types; an example is the `file-info` record whose various fields describe the size, protection, last date of modification, and other pertinent data for a particular file. These record types are described in this manual using the `define-record` notation, which looks like the following:

```
(define-record ship
  x
  y
  (size 100))
```

This form defines a *ship* record, with three fields: its x and y coordinates, and its size. The values of the *x* and *y* fields are specified as parameters to the ship-building procedure, `(make-ship` *x y*`)`, and the *size* field is initialised to 100. All told, the `define-record` form above defines the following procedures:

Procedure	Definition
(make-ship *x y*)	Create a new *ship* record.
(ship:x *ship*)	Retrieve the *x* field.
(ship:y *ship*)	Retrieve the *y* field.
(ship:size *ship*)	Retrieve the *size* field.
(set-ship:x *ship new-x*)	Assign the *x* field.
(set-ship:y *ship new-y*)	Assign the *x* field.
(set-ship:size *ship new-size*)	Assign the *size* field.
(ship? *object*)	Type predicate.

An implementation of `define-record` is available as a macro for Scheme programmers to define their own record types; the syntax is accessed by opening the package `defrec-package`, which exports the single syntax form `define-record`.

You must open this package to access the form. Scsh does not export a record-definition package by default as there are several from which to choose. Besides the `define-record` macro, which Shivers prefers[1], you might instead wish to employ the notationally-distinct `define-record-type` macro that Jonathan Rees prefers,[2] or the identically named but wholly different `define-record-type` macro that Richard Kelsey prefers.[3] The former can be found in file `rts/jar-defrecord.scm` and package `define-record-types`; the latter can be found in file `big/defrecord.scm` and package `defrecord`.

Alternatively, you may define your own, of course.

[1] He wrote it.
[2] He wrote it.
[3] He wrote it.

1.7 A word about Unix standards

"The wonderful thing about Unix standards is that there are so many to choose from." You may be totally bewildered about the multitude of various standards that exist. Rest assured that this nowhere in this manual will you encounter an attempt to spell it all out for you; you could not read and internalise such a twisted account without bleeding from the nose and ears.

However, you might keep in mind the following simple fact: of all the standards, POSIX, as far as I have been able to determine, is the least common denominator. So when this manual repeatedly refers to POSIX, the point is "the thing we are describing should be portable just about anywhere." Scsh sticks to POSIX when at all possible; it's major departure is symbolic links, which aren't in POSIX (see—it really *is* a least common denominator).

However, just because POSIX is the l.c.d. standard doesn't mean everyone supports all of it. The guerilla PC Unix implementations that have been springing up on the net (*e.g.*, NetBSD, Linux, FreeBSD, and so forth) are only recently coming into compliance with the standard—although they are getting there. We have been able to implement scsh completely on all of these systems, however—the single exception is NeXTSTEP, whose buggy POSIX libraries restricts us to partial support (these lacunae are indicated where relevant in the rest of the manual).[4]

[4]Feel like porting scsh from POSIX to NeXT's BSD API? Send us your fixes; we'll fold them in.

Chapter 2

Process notation

Scsh has a notation for controlling Unix processes that takes the form of s-expressions; this notation can then be embedded inside of standard Scheme code. The basic elements of this notation are *process forms*, *extended process forms*, and *redirections*.

2.1 Extended process forms and i/o redirections

An *extended process form* is a specification of a Unix process to run, in a particular I/O environment:

epf ::= (*pf* $redir_1$ ... $redir_n$)

where *pf* is a process form and the $redir_i$ are redirection specs. A *redirection spec* is one of:

```
(< [fdes] file-name)      ; Open file for read.
(> [fdes] file-name)      ; Open file create/truncate.
(<< [fdes] object)        ; Use object's printed rep.
(>> [fdes] file-name)     ; Open file for append.
(= fdes fdes/port)        ; Dup2
(- fdes/port)             ; Close fdes/port.
stdports                  ; 0,1,2 dup'd from standard ports.
```

The input redirections default to file descriptor 0; the output redirections default to file descriptor 1.

The subforms of a redirection are implicitly backquoted, and symbols stand for their print-names. So `(> ,x)` means "output to the file named by Scheme variable `x`," and `(< /usr/shivers/.login)` means "read from `/usr/shivers/.login`."

Here are two more examples of i/o redirection:

```
(< ,(vector-ref fv i))
(>> 2 /tmp/buf)
```

These two redirections cause the file `fv[i]` to be opened on stdin, and `/tmp/buf` to be opened for append writes on stderr.

The redirection `(<<` *object*`)` causes input to come from the printed representation of *object*. For example,

```
(<< "The quick brown fox jumped over the lazy dog.")
```

causes reads from stdin to produce the characters of the above string. The object is converted to its printed representation using the `display` procedure, so

```
(<< (A five element list))
```

is the same as

```
(<< "(A five element list)")
```

is the same as

```
(<< ,(reverse '(list element five A))).
```

(Here we use the implicit backquoting feature to compute the list to be printed.)

The redirection `(=` *fdes fdes/port*`)` causes *fdes/port* to be dup'd into file descriptor *fdes*. For example, the redirection

```
(= 2 1)
```

causes stderr to be the same as stdout. *fdes/port* can also be a port, for example:

```
(= 2 ,(current-output-port))
```

causes stderr to be dup'd from the current output port. In this case, it is an error if the port is not a file port (*e.g.*, a string port). More complex redirections can be accomplished using the `begin` process form, discussed below, which gives the programmer full control of i/o redirection from Scheme.

2.1.1 Port and file descriptor sync

It's important to remember that rebinding Scheme's current I/O ports (*e.g.*, using `call-with-input-file` to rebind the value of `(current-input-port)`) does *not* automatically "rebind" the file referenced by the Unix stdio file descriptors 0, 1, and 2. This is impossible to do in general, since some Scheme ports are not representable as Unix file descriptors. For example, many Scheme implementations provide "string ports," that is, ports that collect characters sent to them into memory buffers. The accumulated string can later be retrieved from the port as a string. If a user were to bind `(current-output-port)` to such a port, it would be impossible to associate file descriptor 1 with this port, as it cannot be represented in Unix. So, if the user subsequently forked off some other program as a subprocess, that program would of course not see the Scheme string port as its standard output.

To keep stdio synced with the values of Scheme's current i/o ports, use the special redirection `stdports`. This causes 0, 1, 2 to be redirected from the current Scheme standard ports. It is equivalent to the three redirections:

```
(= 0 ,(current-input-port))
(= 1 ,(current-output-port))
(= 2 ,(error-output-port))
```

The redirections are done in the indicated order. This will cause an error if one of the current i/o ports isn't a Unix port (*e.g.*, if one is a string port). This Scheme/Unix i/o synchronisation can also be had in Scheme code (as opposed to a redirection spec) with the `(stdports->stdio)` procedure.

2.2 Process forms

A *process form* specifies a computation to perform as an independent Unix process. It can be one of the following:

```
(begin . scheme-code)              ; Run scheme-code in a fork.
(| pf_1 ... pf_n)                  ; Simple pipeline
(|+ connect-list pf_1 ... pf_n)    ; Complex pipeline
(epf . epf)                        ; An extended process form.
(prog arg_1 ... arg_n)             ; Default: exec the program.
```

The default case `(`*prog* arg_1 `...` arg_n`)` is also implicitly backquoted. That is, it is equivalent to:

```
(begin (apply exec-path `(prog arg_1 ... arg_n)))
```

`Exec-path` is the version of the `exec()` system call that uses scsh's path list to search for an executable. The program and the arguments must be either strings, symbols, or integers. Symbols and integers are coerced to strings. A symbol's print-name is used. Integers are converted to strings in base 10. Using symbols instead of strings is convenient, since it suppresses the clutter of the surrounding `"..."` quotation marks. To aid this purpose, scsh reads symbols in a case-sensitive manner, so that you can say

```
(more Readme)
```

and get the right file.

A *connect-list* is a specification of how two processes are to be wired together by pipes. It has the form `((`*from*$_1$ *from*$_2$ `...` *to*`) ...)` and is implicitly backquoted. For example,

```
(|+ ((1 2 0) (3 1)) pf1 pf2)
```

runs pf_1 and pf_2. The first clause `(1 2 0)` causes pf_1's stdout (1) and stderr (2) to be connected via pipe to pf_2's stdin (0). The second clause `(3 1)` causes pf_1's file descriptor 3 to be connected to pf_2's file descriptor 1.

The `begin` process form does a `stdio->stdports` synchronisation in the child process before executing the body of the form. This guarantees that the `begin` form, like all other process forms, "sees" the effects of any associated I/O redirections.

Note that R4RS does not specify whether or not `|` and `|+` are readable symbols. Scsh does.

2.3 Using extended process forms in Scheme

Process forms and extended process forms are *not* Scheme. They are a different notation for expressing computation that, like Scheme, is based upon s-expressions. Extended process forms are used in Scheme programs by embedding them inside special Scheme forms. There are three basic Scheme forms that use extended process forms: `exec-epf`, `&`, and `run`.

`(exec-epf` . *epf*`)` $\longrightarrow$ *no return value* — syntax
`(&` . *epf*`)` $\longrightarrow$ *proc* — syntax
`(run` . *epf*`)` $\longrightarrow$ *proc* — syntax

The `(exec-epf .` *epf*`)` form nukes the current process: it establishes the i/o redirections and then overlays the current process with the requested computation.

The `(& .` *epf*`)` form is similar, except that the process is forked off in background. The form returns the subprocess' process object.

The `(run .` *epf*`)` form runs the process in foreground: after forking off the computation, it waits for the subprocess to exit, and returns its exit status.

These special forms are macros that expand into the equivalent series of system calls. The definition of the `exec-epf` macro is non-trivial, as it produces the code to handle i/o redirections and set up pipelines. However, the definitions of the `&` and `run` macros are very simple:

```
(& . epf)    ≡  (fork (λ () (exec-epf . epf)))
(run . epf)  ≡  (wait (& . epf))
```

2.3.1 Procedures and special forms

It is a general design principle in scsh that all functionality made available through special syntax is also available in a straightforward procedural form. So there are procedural equivalents for all of the process notation. In this way, the programmer is not restricted by the particular details of the syntax. Here are some of the syntax/procedure equivalents:

Notation	Procedure
`|`	`fork/pipe`
`|+`	`fork/pipe+`
`exec-epf`	`exec-path`
redirection	`open`, `dup`
`&`	`fork`
`run`	`wait` + `fork`

Having a solid procedural foundation also allows for general notational experimentation using Scheme's macros. For example, the programmer can build his own pipeline notation on top of the `fork` and `fork/pipe` procedures. Chapter 3 gives the full story on all the procedures in the syscall library.

2.3.2 Interfacing process output to Scheme

There is a family of procedures and special forms that can be used to capture the output of processes as Scheme data.

(`run/port` . *epf*) $\longrightarrow$ *port* syntax
(`run/file` . *epf*) $\longrightarrow$ *string* syntax
(`run/string` . *epf*) $\longrightarrow$ *string* syntax
(`run/strings` . *epf*) $\longrightarrow$ *string list* syntax
(`run/sexp` . *epf*) $\longrightarrow$ *object* syntax
(`run/sexps` . *epf*) $\longrightarrow$ *list* syntax

These forms all fork off subprocesses, collecting the process' output to stdout in some form or another. The subprocess runs with file descriptor 1 and the current output port bound to a pipe.

`run/port`	Value is a port open on process's stdout. Returns immediately after forking child.
`run/file`	Value is name of a temp file containing process's output. Returns when process exits.
`run/string`	Value is a string containing process' output. Returns when eof read.
`run/strings`	Splits process' output into a list of newline-delimited strings. Returns when eof read.
`run/sexp`	Reads a single object from process' stdout with `read`. Returns as soon as the read completes.
`run/sexps`	Repeatedly reads objects from process' stdout with `read`. Returns accumulated list upon eof.

The delimiting newlines are not included in the strings returned by `run/strings`.

These special forms just expand into calls to the following analogous procedures.

`(run/port*` *thunk*`)` ⟶ *port* procedure
`(run/file*` *thunk*`)` ⟶ *string* procedure
`(run/string*` *thunk*`)` ⟶ *string* procedure
`(run/strings*` *thunk*`)` ⟶ *string list* procedure
`(run/sexp*` *thunk*`)` ⟶ *object* procedure
`(run/sexps*` *thunk*`)` ⟶ *object list* procedure

For example, `(run/port .` *epf*`)` expands into

`(run/port* (`λ `() (exec-epf .` *epf*`)))`.

The following procedures are also of utility for generally parsing input streams in scsh:

`(port->string` *port*`)` ⟶ *string* procedure
`(port->sexp-list` *port*`)` ⟶ *list* procedure
`(port->string-list` *port*`)` ⟶ *string list* procedure
`(port->list` *reader port*`)` ⟶ *list* procedure

`Port->string` reads the port until eof, then returns the accumulated string. `Port->sexp-list` repeatedly reads data from the port until eof, then returns the accumulated list of items. `Port->string-list` repeatedly reads newline-terminated strings from the port until eof, then returns the accumulated list of strings. The delimiting newlines are not part of the returned strings. `Port->list` generalises these two procedures. It uses *reader* to repeatedly read objects from a port. It accumulates these objects into a list, which is returned upon eof. The `port->string-list` and `port->sexp-list` procedures are trivial to define, being merely `port->list` curried with the appropriate parsers:

```
(port->string-list port) ≡ (port->list read-line port)
(port->sexp-list   port) ≡ (port->list read port)
```

The following compositions also hold:

```
run/string*   ≡  port->string      ∘ run/port*
run/strings*  ≡  port->string-list ∘ run/port*
run/sexp*     ≡  read              ∘ run/port*
run/sexps*    ≡  port->sexp-list   ∘ run/port*
```

(`port-fold` *port reader op . seeds*) ⟶ *object** procedure

This procedure can be used to perform a variety of iterative operations over an input stream. It repeatedly uses *reader* to read an object from *port*. If the first read returns eof, then the entire `port-fold` operation returns the seeds as multiple values. If the first read operation returns some other value v, then *op* is applied to v and the seeds: (*op* v . *seeds*). This should return a new set of seed values, and the reduction then loops, reading a new value from the port, and so forth. (If multiple seed values are used, then *op* must return multiple values.)

For example, (`port->list` *reader port*) could be defined as

```
(reverse (port-fold port reader cons '()))
```

An imperative way to look at `port-fold` is to say that it abstracts the idea of a loop over a stream of values read from some port, where the seed values express the loop state.

> *Remark:* This procedure was formerly named `reduce-port` . The old binding is still provided, but is deprecated and will probably vanish in a future release.

2.4 More complex process operations

The procedures and special forms in the previous section provide for the common case, where the programmer is only interested in the output of the process. These special forms and procedures provide more complicated facilities for manipulating processes.

2.4.1 Pids and ports together

(run/port+proc . *epf*) ⟶ *[port proc]* syntax
(run/port+proc* *thunk*) ⟶ *[port proc]* procedure

This special form and its analogous procedure can be used if the programmer also wishes access to the process' pid, exit status, or other information. They both fork off a subprocess, returning two values: a port open on the process' stdout (and current output port), and the subprocess's process object. A process object encapsulates the subprocess' process id and exit code; it is the value passed to the `wait` system call.

For example, to uncompress a tech report, reading the uncompressed data into scsh, and also be able to track the exit status of the decompression process, use the following:

```
(receive (port child) (run/port+proc (zcat tr91-145.tex.Z))
  (let* ((paper (port->string port))
         (status (wait child)))
    ...use paper, status, and child here...))
```

Note that you must *first* do the `port->string` and *then* do the wait—the other way around may lock up when the zcat fills up its output pipe buffer.

2.4.2 Multiple stream capture

Occasionally, the programmer may want to capture multiple distinct output streams from a process. For instance, he may wish to read the stdout and stderr streams into two distinct strings. This is accomplished with the `run/collecting` form and its analogous procedure, `run/collecting*`.

(run/collecting *fds . epf*) ⟶ *[port...]* syntax
(run/collecting* *fds thunk*) ⟶ *[port...]* procedure

`Run/collecting` and `run/collecting*` run processes that produce multiple output streams and return ports open on these streams. To avoid issues of deadlock, `run/collecting` doesn't use pipes. Instead, it first runs the process with output to temp files, then returns ports open on the temp files. For example,

```
(run/collecting (1 2) (ls))
```

runs `ls` with stdout (fd 1) and stderr (fd 2) redirected to temporary files. When the `ls` is done, `run/collecting` returns three values: the `ls` process' exit status, and two ports open on the temporary files. The files are deleted before `run/collecting` returns, so when the ports are closed, they vanish. The `fds` list of file descriptors is implicitly backquoted by the special-form version.

For example, if Kaiming has his mailbox protected, then

```
(receive (status out err)
         (run/collecting (1 2) (cat /usr/kmshea/mbox))
  (list status (port->string out) (port->string err)))
```

might produce the list

```
(256 "" "cat: /usr/kmshea/mbox: Permission denied")
```

What is the deadlock hazard that causes `run/collecting` to use temp files? Processes with multiple output streams can lock up if they use pipes to communicate with Scheme i/o readers. For example, suppose some Unix program `myprog` does the following:

1. First, outputs a single "`(`" to stderr.
2. Then, outputs a megabyte of data to stdout.
3. Finally, outputs a single "`)`" to stderr, and exits.

Our scsh programmer decides to run `myprog` with stdout and stderr redirected *via Unix pipes* to the ports `port1` and `port2`, respectively. He gets into trouble when he subsequently says `(read port2)`. The Scheme `read` routine reads the open paren, and then hangs in a `read()` system call trying to read a matching close paren. But before `myprog` sends the close paren down the stderr pipe, it first tries to write a megabyte of data to the stdout pipe. However, Scheme is not reading that pipe—it's stuck waiting for input on stderr. So the stdout pipe quickly fills up, and `myprog` hangs, waiting for the pipe to drain. The `myprog` child is stuck in a stdout/`port1` write; the Scheme parent is stuck in a stderr/`port2` read. Deadlock.

Here's a concrete example that does exactly the above:

```
(receive (status port1 port2)
         (run/collecting (1 2)
             (begin
               ;; Write an open paren to stderr.
               (run (echo "(") (= 1 2))
               ;; Copy a lot of stuff to stdout.
               (run (cat /usr/dict/words))
               ;; Write a close paren to stderr.
               (run (echo ")") (= 1 2))))

   ;; OK. Here, I have a port PORT1 built over a pipe
   ;; connected to the BEGIN subproc's stdout, and
   ;; PORT2 built over a pipe connected to the BEGIN
   ;; subproc's stderr.
   (read port2) ; Should return the empty list.
   (port->string port1)) ; Should return a big string.
```

In order to avoid this problem, `run/collecting` and `run/collecting*` first run the child process to completion, buffering all the output streams in temp files (using the `temp-file-channel` procedure, see below). When the child process exits, ports open on the buffered output are returned. This approach has two disadvantages over using pipes:

- The total output from the child output is temporarily written to the disk before returning from `run/collecting`. If this output is some large intermediate result, the disk could fill up.
- The child producer and Scheme consumer are serialised; there is no concurrency overlap in their execution.

However, it remains a simple solution that avoids deadlock. More sophisticated solutions can easily be programmed up as needed—`run/collecting*` itself is only 12 lines of simple code.

See `temp-file-channel` for more information on creating temp files as communication channels.

2.5 Conditional process sequencing forms

These forms allow conditional execution of a sequence of processes.

`(||` *pf_1 ... pfn*`)` $\longrightarrow$ *boolean* syntax

Run each proc until one completes successfully (*i.e.*, exit status zero). Return true if some proc completes successfully; otherwise `#f`.

(`&&` $pf_1 \ldots pfn$) $\longrightarrow$ *boolean* syntax

Run each proc until one fails (*i.e.*, exit status non-zero). Return true if all procs complete successfully; otherwise `#f`.

2.6 Process filters

These procedures are useful for forking off processes to filter text streams.

(`char-filter` *filter*) $\longrightarrow$ *procedure* procedure

The *filter* argument is a character→character procedure. Returns a procedure that when called, repeatedly reads a character from the current input port, applies *filter* to the character, and writes the result to the current output port. The procedure returns upon reaching eof on the input port.

For example, to downcase a stream of text in a spell-checking pipeline, instead of using the Unix `tr A-Z a-z` command, we can say:

```
(run (| (delatex)
        (begin ((char-filter char-downcase))) ; tr A-Z a-z
        (spell)
        (sort)
        (uniq))
     (< scsh.tex)
     (> spell-errors.txt))
```

(`string-filter` *filter [buflen]*) $\longrightarrow$ *procedure* procedure

The *filter* argument is a string→string procedure. Returns a procedure that when called, repeatedly reads a string from the current input port, applies *filter* to the string, and writes the result to the current output port. The procedure returns upon reaching eof on the input port.

The optional *buflen* argument controls the number of characters each internal read operation requests; this means that *filter* will never be applied to a string longer than *buflen* chars. The default *buflen* value is 1024.

Chapter 3

System Calls

Scsh provides (almost) complete access to the basic Unix kernel services: processes, files, signals and so forth. These procedures comprise a Scheme binding for POSIX, with a few of the more standard extensions thrown in (*e.g.*, symbolic links, `fchown`, `fstat`, sockets).

3.1 Errors

Scsh syscalls never return error codes, and do not use a global `errno` variable to report errors. Errors are consistently reported by raising exceptions. This frees up the procedures to return useful values, and allows the programmer to assume that *if a syscall returns, it succeeded.* This greatly simplifies the flow of the code from the programmer's point of view.

Since Scheme does not yet have a standard exception system, the scsh definition remains somewhat vague on the actual form of exceptions and exception handlers. When a standard exception system is defined, scsh will move to it. For now, scsh uses the Scheme 48 exception system, with a simple sugaring on top to hide the details in the common case.

System call error exceptions contain the Unix `errno` code reported by the system call. Unlike C, the `errno` value is a part of the exception packet, it is *not* accessed through a global variable.

For reference purposes, the Unix `errno` numbers are bound to the variables `errno/perm`, `errno/noent`, *etc.* System calls never return `error/intr`—they automatically retry.

(`errno-error` *errno syscall . data*) $\longrightarrow$ *no return value* procedure

Raises a Unix error exception for Unix error number *errno*. The *syscall* and *data* arguments are packaged up in the exception packet passed to the exception handler.

`(with-errno-handler*` *handler thunk*`)` $\longrightarrow$ *value(s) of thunk* procedure
`(with-errno-handler` *handler-spec . body*`)` $\longrightarrow$ *value of body* syntax

Unix syscalls raise error exceptions by calling `errno-error`. Programs can use `with-errno-handler*` to establish handlers for these exceptions.

If a Unix error arises while *thunk* is executing, *handler* is called on two arguments:

(*handler errno packet*)

packet is a list of the form

$$packet = (errno\text{-}msg \;\; syscall \;\; . \;\; data),$$

where *errno-msg* is the standard Unix error message for the error, *syscall* is the procedure that generated the error, and *data* is a list of information generated by the error, which varies from syscall to syscall.

If *handler* returns, the handler search continues upwards. *Handler* can acquire the exception by invoking a saved continuation. This procedure can be sugared over with the following syntax:

```
(with-errno-handler
    ((errno packet) clause ...)
  body1
  body2
  ...)
```

This form executes the body forms with a particular errno handler installed. When an errno error is raised, the handler search machinery will bind variable *errno* to the error's integer code, and variable *packet* to the error's auxiliary data packet. Then, the clauses will be checked for a match. The first clause that matches is executed, and its value is the value of the entire `with-errno-handler` form. If no clause matches, the handler search continues.

Error clauses have two forms

```
((errno ...) body ...)
(else body ...)
```

In the first type of clause, the *errno* forms are integer expressions. They are evaluated and compared to the error's errno value. An `else` clause matches any errno value. Note that the *errno* and *data* variables are lexically visible to the error clauses.

Example:

```
(with-errno-handler
    ((errno packet) ; Only handle 3 particular errors.
     ((errno/wouldblock errno/again)
      (loop))
     ((errno/acces)
      (format #t "Not allowed access!")
      #f))

  (foo frobbotz)
  (blatz garglemumph))
```

It is not defined what dynamic context the handler executes in, so fluid variables cannot reliably be referenced.

Note that Scsh system calls always retry when interrupted, so that the `errno/intr` exception is never raised. If the programmer wishes to abort a system call on an interrupt, he should have the interrupt handler explicitly raise an exception or invoke a stored continuation to throw out of the system call.

3.1.1 Interactive mode and error handling

Scsh runs in two modes: interactive and script mode. It starts up in interactive mode if the scsh interpreter is started up with no script argument. Otherwise, scsh starts up in script mode. The mode determines whether scsh prints prompts in between reading and evaluating forms, and it affects the default error handler. In interactive mode, the default error handler will report the error, and generate an interactive breakpoint so that the user can interact with the system to examine, fix, or dismiss from the error. In script mode, the default error handler causes the scsh process to exit.

When scsh forks a child with `(fork)`, the child resets to script mode. This can be overridden if the programmer wishes.

3.2 I/O

3.2.1 Standard R4RS I/O procedures

In scsh, most standard R4RS `i/o` operations (such as `display` or `read-char`) work on both integer file descriptors and Scheme ports. When doing `i/o` with a file descriptor, the `i/o` operation is done directly on the file, bypassing any buffered data that may have accumulated in an associated port. Note that character-at-a-time operations such as `read-char` are likely to be quite slow when performed directly upon file descriptors.

The standard R4RS procedures `read-char`, `char-ready?`, `write`, `display`, `newline`, and `write-char` are all generic, accepting integer file descriptor arguments as well as ports. Scsh also mandates the availability of `format`, and further requires `format` to accept file descriptor arguments as well as ports.

The procedures `peek-char` and `read` do *not* accept file descriptor arguments, since these functions require the ability to read ahead in the input stream, a feature not supported by Unix I/O.

3.2.2 Port manipulation and standard ports

`(close-after` *port consumer*`)` $\longrightarrow$ *value(s) of consumer* procedure

Returns `(`*consumer port*`)`, but closes the port on return. No dynamic-wind magic.

> *Remark:* Is there a less-awkward name?

`(error-output-port)` $\longrightarrow$ *port* procedure

This procedure is analogous to `current-output-port`, but produces a port used for error messages—the scsh equivalent of stderr.

`(with-current-input-port*` *port thunk*`)` $\longrightarrow$ *value(s) of thunk* procedure
`(with-current-output-port*` *port thunk*`)` $\longrightarrow$ *value(s) of thunk* procedure
`(with-error-output-port*` *port thunk*`)` $\longrightarrow$ *value(s) of thunk* procedure

These procedures install *port* as the current input, current output, and error output port, respectively, for the duration of a call to *thunk*.

`(with-current-input-port` *port . body*`)` $\longrightarrow$ *value(s) of body* syntax
`(with-current-output-port` *port . body*`)` $\longrightarrow$ *value(s) of body* syntax
`(with-error-output-port` *port . body*`)` $\longrightarrow$ *value(s) of body* syntax

These special forms are simply syntactic sugar for the `with-current-input-port*` procedure and friends.

`(set-current-input-port! port)` ⟶ *undefined* procedure
`(set-current-output-port! port)` ⟶ *undefined* procedure
`(set-error-output-port! port)` ⟶ *undefined* procedure

These procedures alter the dynamic binding of the current I/O port procedures to new values.

`(close fd/port)` ⟶ *boolean* procedure

Close the port or file descriptor.

If *fd/port* is a file descriptor, and it has a port allocated to it, the port is shifted to a new file descriptor created with `(dup fd/port)` before closing `fd/port`. The port then has its revealed count set to zero. This reflects the design criteria that ports are not associated with file descriptors, but with open files.

To close a file descriptor, and any associated port it might have, you must instead say one of (as appropriate):

```
(close (fdes->inport  fd))
(close (fdes->outport fd))
```

The procedure returns true if it closed an open port. If the port was already closed, it returns false; this is not an error.

`(stdports->stdio)` ⟶ *undefined* procedure
`(stdio->stdports)` ⟶ *undefined* procedure

These two procedures are used to synchronise Unix' standard I/O file descriptors and Scheme's current I/O ports.

`(stdports->stdio)` causes the standard I/O file descriptors (0, 1, and 2) to take their values from the current I/O ports. It is exactly equivalent to the series of redirections:[1]

```
(dup (current-input-port)  0)
(dup (current-output-port) 1)
(dup (error-output-port)   2)
```

`stdio->stdports` causes the bindings of the current I/O ports to be changed to ports constructed over the standard I/O file descriptors. It is exactly equivalent to the series of assignments

```
(set-current-input-port!  (fdes->inport  0))
(set-current-output-port! (fdes->outport 1))
(set-error-output-port!   (fdes->outport 2))
```

[1] Why not `move->fdes`? Because the current output port and error port might be the same port.

However, you are more likely to find the dynamic-extent variant, `with-stdio-ports*`, below, to be of use in general programming.

(`with-stdio-ports*` *thunk*) $\longrightarrow$ *value(s) of thunk* procedure
(`with-stdio-ports` *body ...*) $\longrightarrow$ *value(s) of body* syntax

`with-stdio-ports*` binds the standard ports `(current-input-port)`, `(current-output-port)`, and `(error-output-port)` to be ports on file descriptors 0, 1, 2, and then calls *thunk*. It is equivalent to:

```
(with-current-input-port (fdes->inport 0)
  (with-current-output-port (fdes->inport 1)
    (with-error-output-port (fdes->outport 2)
      (thunk))))
```

The `with-stdio-ports` special form is merely syntactic sugar.

3.2.3 String ports

Scheme 48 has string ports, which you can use. Scsh has not committed to the particular interface or names that Scheme 48 uses, so be warned that the interface described herein may be liable to change.

(`make-string-input-port` *string*) $\longrightarrow$ *port* procedure

Returns a port that reads characters from the supplied string.

(`make-string-output-port`) $\longrightarrow$ *port* procedure
(`string-output-port-output` *port*) $\longrightarrow$ *string* procedure

A string output port is a port that collects the characters given to it into a string. The accumulated string is retrieved by applying `string-output-port-output` to the port.

(`call-with-string-output-port` *procedure*) $\longrightarrow$ *string* procedure

The *procedure* value is called on a port. When it returns, `call-with-string-output-port` returns a string containing the characters that were written to that port during the execution of *procedure*.

3.2.4 Revealed ports and file descriptors

The material in this section and the following one is not critical for most applications. You may safely skim or completely skip this section on a first reading.

Dealing with Unix file descriptors in a Scheme environment is difficult. In Unix, open files are part of the process environment, and are referenced by small integers called *file descriptors*. Open file descriptors are the fundamental way i/o redirections are passed to subprocesses, since file descriptors are preserved across fork's and exec's.

Scheme, on the other hand, uses ports for specifying i/o sources. Ports are garbage-collected Scheme objects, not integers. Ports can be garbage collected; when a port is collected, it is also closed. Because file descriptors are just integers, it's impossible to garbage collect them—you wouldn't be able to close file descriptor 3 unless there were no 3's in the system, and you could further prove that your program would never again compute a 3. This is difficult at best.

If a Scheme program only used Scheme ports, and never actually used file descriptors, this would not be a problem. But Scheme code must descend to the file descriptor level in at least two circumstances:

- when interfacing to foreign code
- when interfacing to a subprocess.

This causes a problem. Suppose we have a Scheme port constructed on top of file descriptor 2. We intend to fork off a program that will inherit this file descriptor. If we drop references to the port, the garbage collector may prematurely close file 2 before we fork the subprocess. The interface described below is intended to fix this and other problems arising from the mismatch between ports and file descriptors.

The Scheme kernel maintains a port table that maps a file descriptor to the Scheme port allocated for it (or, `#f` if there is no port allocated for this file descriptor). This is used to ensure that there is at most one open port for each open file descriptor.

The port data structure for file ports has two fields besides the descriptor: *revealed* and *closed?*. When a file port is closed with `(close port)`, the port's file descriptor is closed, its entry in the port table is cleared, and the port's *closed?* field is set to true.

When a file descriptor is closed with `(close fdes)`, any associated port is shifted to a new file descriptor created with `(dup fdes)`. The port has its revealed count reset to zero (and hence becomes eligible for closing on GC). See discussion below. To really put a stake through a descriptor's heart without waiting for associated ports to be GC'd, you must say one of

```
(close (fdes->inport fdes))
(close (fdes->output fdes))
```

The *revealed* field is an aid to garbage collection. It is an integer semaphore. If it is zero, the port's file descriptor can be closed when the port is collected.

Essentially, the *revealed* field reflects whether or not the port's file descriptor has escaped to the Scheme user. If the Scheme user doesn't know what file descriptor is associated with a given port, then he can't possibly retain an "integer handle" on the port after dropping pointers to the port itself, so the garbage collector is free to close the file.

Ports allocated with `open-output-file` and `open-input-file` are unrevealed ports—*i.e.*, *revealed* is initialised to 0. No one knows the port's file descriptor, so the file descriptor can be closed when the port is collected.

The functions `fdes->output-port`, `fdes->input-port`, `port->fdes` are used to shift back and forth between file descriptors and ports. When `port->fdes` reveals a port's file descriptor, it increments the port's *revealed* field. When the user is through with the file descriptor, he can call (`release-port-handle` *port*), which decrements the count. The function (`call/fdes fd/port` *proc*) automates this protocol. `call/fdes` uses `dynamic-wind` to enforce the protocol. If *proc* throws out of the `call/fdes` application, the unwind handler releases the descriptor handle; if the user subsequently tries to throw *back* into *proc*'s context, the wind handler raises an error. When the user maps a file descriptor to a port with `fdes->outport` or `fdes->inport`, the port has its revealed field incremented.

Not all file descriptors are created by requests to make ports. Some are inherited on process invocation via `exec(2)`, and are simply part of the global environment. Subprocesses may depend upon them, so if a port is later allocated for these file descriptors, is should be considered as a revealed port. For example, when the Scheme shell's process starts up, it opens ports on file descriptors 0, 1, and 2 for the initial values of `(current-input-port)`, `(current-output-port)`, and `(error-output-port)`. These ports are initialised with *revealed* set to 1, so that stdin, stdout, and stderr are not closed even if the user drops the port.

Unrevealed file ports have the nice property that they can be closed when all pointers to the port are dropped. This can happen during gc, or at an `exec()`—since all memory is dropped at an `exec()`. No one knows the file descriptor associated with the port, so the exec'd process certainly can't refer to it.

This facility preserves the transparent close-on-collect property for file ports that are used in straightforward ways, yet allows access to the underlying Unix substrate without interference from the garbage collector. This is critical, since shell programming absolutely requires access to the Unix file descriptors, as their numerical values are a critical part of the process interface.

A port's underlying file descriptor can be shifted around with `dup(2)` when convenient. That is, the actual file descriptor on top of which a port is constructed can be shifted around underneath the port by the scsh kernel when necessary. This is important, because when the user is setting up file descriptors prior to a `exec(2)`, he may explicitly use a file descriptor that has already been allocated to some port. In this case, the scsh kernel just shifts the port's file descriptor to some new location with `dup`, freeing up its old descriptor. This prevents errors from happening in the following scenario. Suppose we have a file open on port `f`. Now we want to run a program that reads input on file 0, writes output to file 1, errors to file 2, and logs execution information on file 3. We want to run this program with input from `f`. So we write:

```
(run (/usr/shivers/bin/prog)
     (> 1 output.txt)
     (> 2 error.log)
     (> 3 trace.log)
     (= 0 ,f))
```

Now, suppose by ill chance that, unbeknownst to us, when the operating system opened `f`'s file, it allocated descriptor 3 for it. If we blindly redirect `trace.log` into file descriptor 3, we'll clobber `f`! However, the port-shuffling machinery saves us: when the `run` form tries to dup `trace.log`'s file descriptor to 3, `dup` will notice that file descriptor 3 is already associated with an unrevealed port (*i.e.*, `f`). So, it will first move `f` to some other file descriptor. This keeps `f` alive and well so that it can subsequently be dup'd into descriptor 0 for `prog`'s stdin.

The port-shifting machinery makes the following guarantee: a port is only moved when the underlying file descriptor is closed, either by a `close()` or a `dup2()` operation. Otherwise a port/file-descriptor association is stable.

Under normal circumstances, all this machinery just works behind the scenes to keep things straightened out. The only time the user has to think about it is when he starts accessing file descriptors from ports, which he should almost never have to do. If a user starts asking what file descriptors have been allocated to what ports, he has to take responsibility for managing this information.

3.2.5 Port-mapping machinery

The procedures provided in this section are almost never needed. You may safely skim or completely skip this section on a first reading.

Here are the routines for manipulating ports in scsh. The important points to remember are:

- A file port is associated with an open file, not a particular file descriptor.

- The association between a file port and a particular file descriptor is never changed *except* when the file descriptor is explicitly closed. "Closing" includes being used as the target of a `dup2`, so the set of procedures below that close their targets are `close`, two-argument `dup`, and `move->fdes`. If the target file descriptor of one of these routines has an allocated port, the port will be shifted to another freshly-allocated file descriptor, and marked as unrevealed, thus preserving the port but freeing its old file descriptor.

These rules are what is necessary to "make things work out" with no surprises in the general case.

`(fdes->inport` *fd*`)` $\longrightarrow$ *port* procedure
`(fdes->outport` *fd*`)` $\longrightarrow$ *port* procedure
`(port->fdes` *port*`)` $\longrightarrow$ *fixnum* procedure

These increment the port's revealed count.

`(port-revealed` *port*`)` $\longrightarrow$ *integer or* `#f` procedure

Return the port's revealed count if positive, otherwise `#f`.

`(release-port-handle` *port*`)` $\longrightarrow$ *undefined* procedure

Decrement the port's revealed count.

`(call/fdes` *fd/port consumer*`)` $\longrightarrow$ *value(s) of consumer* procedure

Calls *consumer* on a file descriptor; takes care of revealed bookkeeping. If *fd/port* is a file descriptor, this is just `(`*consumer fd/port*`)`. If *fd/port* is a port, calls *consumer* on its underlying file descriptor. While *consumer* is running, the port's revealed count is incremented.

When `call/fdes` is called with port argument, you are not allowed to throw into *consumer* with a stored continuation, as that would violate the revealed-count bookkeeping.

`(move->fdes` *fd/port target-fd*`)` $\longrightarrow$ *port or fdes* procedure

Maps fd$\rightarrow$fd and port$\rightarrow$port.

If *fd/port* is a file-descriptor not equal to *target-fd*, dup it to *target-fd* and close it. Returns *target-fd*.

If *fd/port* is a port, it is shifted to *target-fd*, by duping its underlying file-descriptor if necessary. *Fd/port*'s original file descriptor is closed (if it was different from *target-fd*). Returns the port. This operation resets *fd/port*'s revealed count to 1.

In all cases when *fd/port* is actually shifted, if there is a port already using *target-fd*, it is first relocated to some other file descriptor.

3.2.6 Unix I/O

(`dup` *fd/port [newfd]*) ⟶ *fd/port* procedure
(`dup->inport` *fd/port [newfd]*) ⟶ *port* procedure
(`dup->outport` *fd/port [newfd]*) ⟶ *port* procedure
(`dup->fdes` *fd/port [newfd]*) ⟶ *fd* procedure

These procedures provide the functionality of C's `dup()` and `dup2()`. The different routines return different types of values: `dup->inport`, `dup->outport`, and `dup->fdes` return input ports, output ports, and integer file descriptors, respectively. `dup`'s return value depends on on the type of *fd/port*—it maps fd→fd and port→port.

These procedures use the Unix `dup()` syscall to replicate the file descriptor or file port *fd/port*. If a *newfd* file descriptor is given, it is used as the target of the dup operation, *i.e.*, the operation is a `dup2()`. In this case, procedures that return a port (such as `dup->inport`) will return one with the revealed count set to one. For example, `(dup (current-input-port) 5)` produces a new port with underlying file descriptor 5, whose revealed count is 1. If *newfd* is not specified, then the operating system chooses the file descriptor, and any returned port is marked as unrevealed.

If the *newfd* target is given, and some port is already using that file descriptor, the port is first quietly shifted (with another `dup`) to some other file descriptor (zeroing its revealed count).

Since Scheme doesn't provide read/write ports, `dup->inport` and `dup->outport` can be useful for getting an output version of an input port, or *vice versa*. For example, if `p` is an input port open on a tty, and we would like to do output to that tty, we can simply use `(dup->outport p)` to produce an equivalent output port for the tty.

(`seek` *fd/port offset [whence]*) ⟶ *integer* procedure

Reposition the I/O cursor for a file descriptor or port. *whence* is one of {`seek/set`, `seek/delta`, `seek/end`}, and defaults to `seek/set`. If `seek/set`, then *offset* is an absolute index into the file; if `seek/delta`, then *offset* is a relative offset from the current I/O cursor; if `seek/end`, then *offset* is a relative offset from the end of file. The *fd/port* argument may be a port or an integer file descriptor. Not all such values are seekable; this is dependent on the OS implementation. The return value is the resulting position of the I/O cursor in the I/O stream.

> *Oops:* The current implementation doesn't handle *offset* arguments that are not immediate integers (*i.e.*, representable in 30 bits).

(`tell` *fd/port*) ⟶ *integer* procedure

Returns the position of the I/O cursor in the the I/O stream. Not all file descriptors or ports support cursor-reporting; this is dependent on the OS implementation.

`(open-file` *fname flags [perms]*`)` $\longrightarrow$ *port* procedure

Perms defaults to `#o666`. *Flags* is an integer bitmask, composed by or'ing together constants listed in table 3.1 (page 33). You must use exactly one of the `open/read`, `open/write`, or `open/read+write` flags. The returned port is an input port if the *flags* permit it, otherwise an output port. R4RS/Scheme 48/scsh do not have input/output ports, so it's one or the other. This should be fixed. (You can hack simultaneous `i/o` on a file by opening it `r/w`, taking the result input port, and duping it to an output port with `dup->outport`.)

`(open-input-file` *fname [flags]*`)` $\longrightarrow$ *port* procedure
`(open-output-file` *fname [flags perms]*`)` $\longrightarrow$ *port* procedure

These are equivalent to `open-file`, after first setting the read/write bits of the *flags* argument to `open/read` or `open/write`, respectively. *Flags* defaults to zero for `open-input-file`, and

```
(bitwise-ior open/create open/truncate)
```

for `open-output-file`. These defaults make the procedures backwards-compatible with their unary R4RS definitions.

`(open-fdes` *fname flags [perms]*`)` $\longrightarrow$ *integer* procedure

Returns a file descriptor.

`(fdes-flags` *fd/port*`)` $\longrightarrow$ *integer* procedure
`(set-fdes-flags` *fd/port integer*`)` $\longrightarrow$ *undefined* procedure

These procedures allow reading and writing of an open file's flags. The only such flag defined by POSIX is `fdflags/close-on-exec`; your Unix implementation may provide others.

These procedures should not be particularly useful to the programmer, as the scsh runtime already provides automatic control of the close-on-exec property. Unrevealed ports always have their file descriptors marked close-on-exec, as they can be closed when the scsh process execs a new program. Whenever the user reveals or unreveals a port's file descriptor, the runtime automatically sets or clears the flag for the programmer. Programmers that manipulate this flag should be aware of these extra, automatic operations.

	Allowed operations	Status flag

Open+Get+Set	These flags can be used in `open-file`, `fdes-status`, and `set-fdes-status` calls.	`open/append`
		`open/non-blocking`
		`open/async` (Non-POSIX)
		`open/fsync` (Non-POSIX)

Open+Get	These flags can be used in `open-file` and `fdes-status` calls, but are ignored by `set-fdes-status`.	`open/read`
		`open/write`
		`open/read+write`
		`open/access-mask`

Open	These flags are only relevant in `open-file` calls; they are ignored by `fdes-status` and `set-fdes-status` calls.	`open/create`
		`open/exclusive`
		`open/no-control-tty`
		`open/truncate`

Table 3.1: Status flags for `open-file`, `fdes-status` and `set-fdes-status`. Only POSIX flags are guaranteed to be present; your operating system may define others. The `open/access-mask` value is not an actual flag, but a bit mask used to select the field for the `open/read`, `open/write` and `open/read+write` bits.

`(fdes-status` *fd/port*`)` $\longrightarrow$ *integer* procedure
`(set-fdes-status` *fd/port integer*`)` $\longrightarrow$ *undefined* procedure

These procedures allow reading and writing of an open file's status flags (table 3.1).

Note that this file-descriptor state is shared between file descriptors created by `dup`—if you create port *b* by applying `dup` to port *a*, and change *b*'s status flags, you will also have changed *a*'s status flags.

`(pipe)` $\longrightarrow$ *[rport wport]* procedure

Returns two ports, the read and write end-points of a Unix pipe.

`(read-string` *nbytes [fd/port]*`)` $\longrightarrow$ *string or* `#f` procedure
`(read-string!` *str [fd/port start end]*`)` $\longrightarrow$ *nread or* `#f` procedure

These calls read exactly as much data as you requested, unless there is not enough data (eof). `read-string!` reads the data into string *str* at the indices in the half-open interval $[start, end)$; the default interval is the whole string: $start = 0$ and $end =$ `(string-length` *string*`)`. They will persistently retry on partial reads and when interrupted until (1) error,

(2) eof, or (3) the input request is completely satisfied. Partial reads can occur when reading from an intermittent source, such as a pipe or tty.

`read-string` returns the string read; `read-string!` returns the number of characters read. They both return false at eof. A request to read zero bytes returns immediately, with no eof check.

The values of *start* and *end* must specify a well-defined interval in *str*, *i.e.*, $0 \leq start \leq end \leq$ `(string-length` *str*`)`.

Any partially-read data is included in the error exception packet. Error returns on non-blocking input are considered an error.

`(read-string/partial` *nbytes [fd/port]*`)` $\longrightarrow$ *string or* `#f` procedure
`(read-string!/partial` *str [fd/port start end]*`)` $\longrightarrow$ *nread or* `#f` procedure

These are atomic best-effort/forward-progress calls. Best effort: they may read less than you request if there is a lesser amount of data immediately available (*e.g.*, because you are reading from a pipe or a tty). Forward progress: if no data is immediately available (*e.g.*, empty pipe), they will block. Therefore, if you request an $n > 0$ byte read, while you may not get everything you asked for, you will always get something (barring eof).

There is one case in which the forward-progress guarantee is cancelled: when the programmer explicitly sets the port to non-blocking i/o. In this case, if no data is immediately available, the procedure will not block, but will immediately return a zero-byte read.

`read-string/partial` reads the data into a freshly allocated string, which it returns as its value. `read-string!/partial` reads the data into string *str* at the indices in the half-open interval $[start, end)$; the default interval is the whole string: $start = 0$ and $end =$ `(string-length` *string*`)`. The values of *start* and *end* must specify a well-defined interval in *str*, *i.e.*, $0 \leq start \leq end \leq$ `(string-length` *str*`)`. It returns the number of bytes read.

A request to read zero bytes returns immediatedly, with no eof check.

In sum, there are only three ways you can get a zero-byte read: (1) you request one, (2) you turn on non-blocking i/o, or (3) you try to read at eof.

These are the routines to use for non-blocking input. They are also useful when you wish to efficiently process data in large blocks, and your algorithm is insensitive to the block size of any particular read operation.

`(select` *rvec wvec evec [timeout]*`)` $\longrightarrow$ *[rvec' wvec' evec']* procedure
`(select!` *rvec wvec evec [timeout]*`)` $\longrightarrow$ *[nr nw ne]* procedure

The `select` procedure allows a process to block and wait for events on multiple I/O channels. The *rvec* and *evec* arguments are vectors of input ports and integer file descriptors; *wvec* is a vector of output ports and integer file descriptors. The procedure returns three vectors whose elements are subsets of the corresponding arguments. Every element of *rvec′* is ready for input; every element of *wvec′* is ready for output; every element of *evec′* has an exceptional condition pending.

The `select` call will block until at least one of the I/O channels passed to it is ready for operation. The *timeout* value can be used to force the call to time-out after a given number of seconds. It defaults to the special value `#f`, meaning wait indefinitely. A zero value can be used to poll the I/O channels.

If an I/O channel appears more than once in a given vector—perhaps occuring once as a Scheme port, and once as the port's underlying integer file descriptor—only one of these two references may appear in the returned vector. Buffered I/O ports are handled specially—if an input port's buffer is not empty, or an output port's buffer is not yet full, then these ports are immediately considered eligible for I/O without using the actual, primitive `select` system call to check the underlying file descriptor. This works pretty well for buffered input ports, but is a little problematic for buffered output ports.

The `select!` procedure is similar, but indicates the subset of active I/O channels by side-effecting the argument vectors. Non-active I/O channels in the argument vectors are overwritten with `#f` values. The call returns the number of active elements remaining in each vector. As a convenience, the vectors passed in to `select!` are allowed to contain `#f` values as well as integers and ports.

> *Remark:* I have found the `select!` interface to be the more useful of the two. After the system call, it allows you to check a specific I/O channel in constant time.

(`write-string` *string [fd/port start end]*) $\longrightarrow$ *undefined* procedure

This procedure writes all the data requested. If the procedure cannot perform the write with a single kernel call (due to interrupts or partial writes), it will perform multiple write operations until all the data is written or an error has occurred. A non-blocking `i/o` error is considered an error. (Error exception packets for this syscall include the amount of data partially transferred before the error occurred.)

The data written are the characters of *string* in the half-open interval $[start, end)$. The default interval is the whole string: $start = 0$ and $end =$ (`string-length` *string*). The values of *start* and *end* must specify a well-defined interval in *str*, *i.e.*, $0 \leq start \leq end \leq$ (`string-length` *str*). A zero-byte write returns immediately, with no error.

Output to buffered ports: `write-string`'s efforts end as soon as all the data has been placed in the output buffer. Errors and true output may not happen until a later time, of course.

(`write-string/partial` *string [fd/port start end]*) $\longrightarrow$ *nwritten* procedure

This routine is the atomic best-effort/forward-progress analog to `write-string`. It returns the number of bytes written, which may be less than you asked for. Partial writes can occur when (1) we write off the physical end of the media, (2) the write is interrrupted, or (3) the file descriptor is set for non-blocking i/o.

If the file descriptor is not set up for non-blocking i/o, then a successful return from these procedures makes a forward progress guarantee—that is, a partial write took place of at least one byte:

- If we are at the end of physical media, and no write takes place, an error exception is raised. So a return implies we wrote *something*.
- If the call is interrupted after a partial transfer, it returns immediately. But if the call is interrupted before any data transfer, then the write is retried.

If we request a zero-byte write, then the call immediately returns 0. If the file descriptor is set for non-blocking i/o, then the call may return 0 if it was unable to immediately write anything (*e.g.*, full pipe). Barring these two cases, a write either returns $nwritten > 0$, or raises an error exception.

Non-blocking i/o is only available on file descriptors and unbuffered ports. Doing non-blocking i/o to a buffered port is not well-defined, and is an error (the problem is the subsequent flush operation).

3.2.7 Buffered I/O

Scheme 48 ports use buffered I/O—data is transferred to or from the OS in blocks. Scsh provides control of this mechanism: the programmer may force saved-up output data to be transferred to the OS when he chooses, and may also choose which I/O buffering policy to employ for a given port (or turn buffering off completely).

It can be useful to turn I/O buffering off in some cases, for example when an I/O stream is to be shared by multiple subprocesses. For this reason, scsh allocates an unbuffered port for file descriptor 0 at start-up time. Because shells frequently share stdin with subprocesses, if the shell does buffered reads, it might "steal" input intended for a subprocess. For this reason, all shells, including sh, csh, and scsh, read stdin unbuffered. Applications that can tolerate buffered input on stdin can reset `(current-input-port)` to block buffering for higher performance.

`(set-port-buffering` *port policy [size]*`)` $\longrightarrow$ *undefined* procedure

This procedure allows the programmer to assign a particular I/O buffering policy to a port, and to choose the size of the associated buffer. It may only be used on new ports, *i.e.*, before I/O is performed on the port. There are three buffering policies that may be chosen:

`bufpol/block`	General block buffering (general default)
`bufpol/line`	Line buffering (tty default)
`bufpol/none`	Direct I/O—no buffering

The line buffering policy flushes output whenever a newline is output; whenever the buffer is full; or whenever an input is read from stdin. Line buffering is the default for ports open on terminal devices.

The *size* argument requests an I/O buffer of *size* bytes. If not given, a reasonable default is used; if given and zero, buffering is turned off (*i.e.*, *size* $= 0$ for any policy is equivalent to *policy* $=$ `bufpol/none`).

`(force-output` *[fd/port]*`)` $\longrightarrow$ *undefined* procedure

This procedure does nothing when applied to an integer file descriptor or unbuffered port. It flushes buffered output when applied to a buffered port, and raises a write-error exception on error. Returns no value.

`(flush-all-ports)` $\longrightarrow$ *undefined* procedure

This procedure flushes all open output ports with buffered data.

3.2.8 File locking

Scsh provides POSIX advisory file locking. *Advisory* locks are locks that can be checked by user code, but do not affect other I/O operations. For example, if a process has an exclusive lock on a region of a file, other processes will not be able to obtain locks on that region of the file, but they will still be able to read and write the file with no hindrance. Using advisory locks requires cooperation amongst the agents accessing the shared resource.

Remark: Unfortunately, POSIX file locks are associated with actual files, not with associated open file descriptors. Once a process locks a file, using some file descriptor *fd*, the next time *any* file descriptor referencing that file is closed, all associated locks are released. This severely limits the utility of POSIX advisory file locks, and we'd recommend caution when using them. It is not without reason that the FreeBSD man pages refer to POSIX file locking as "completely stupid."

Scsh moves Scheme ports from file descriptor to file descriptor with `dup()` and `close()` as required by the runtime, so it is impossible to keep file locks open across one of these shifts. Hence we can only offer POSIX advisory file locking directly on raw integer file descriptors; regrettably, there are no facilities for locking Scheme ports.

Note that once a Scheme port is revealed in scsh, the runtime will not shift the port around with `dup()` and `close()`. This means the file-locking procedures can then be applied to the port's associated file descriptor.

NeXTSTEP users should also note that even minimalist POSIX file locking is not supported for NFS-mounted files in NeXTSTEP; NeXT claims they will fix this in NS release 4. We'd appreciate hearing from users when and if this happens.

POSIX allows the user to lock a region of a file with either an exclusive or shared lock. Locked regions are described by the *lock-region* record:

```
(define-record lock-region
  exclusive?
  start
  len
  whence
  proc)
```

- The `exclusive?` field is true if the lock is exclusive; false if it is shared.
- The `whence` field is one of the values from the `seek` call: `seek/set`, `seek/delta`, or `seek/end`, and determines the interpretation of the `start` field:
  - If `seek/set`, the `start` value is simply an absolute index into the file.
  - If `seek/delta`, the `start` value is an offset from the file descriptor's current position in the file.
  - If `seek/end`, the `start` value is an offset from the end of the file.

  The region of the file being locked is given by the `start` and `len` fields; if `len` is zero, it means "infinity," that is, the region extends from the starting point through the end of the file, even as the file is extended by subsequent write operations.

- The `proc` field gives the process object for the process holding the region lock, when relevant (see `get-lock-region` below).

(`make-lock-region` *exclusive? start len [whence]*) ⟶ *lock-region* procedure

This procedure makes a lock-region record. The `whence` field defaults to `seek/set`.

(`lock-region` *fdes lock*) ⟶ *undefined* procedure
(`lock-region/no-block` *fdes lock*) ⟶ *boolean* procedure

These procedures lock a region of the file referenced by file descriptor *fdes*. The `lock-region` procedure blocks until the lock is granted; the non-blocking variant returns a boolean indicating whether or not the lock was granted. To take an exclusive (write) lock, you must have the file descriptor open with write access; to take a shared (read) lock, you must have the file descriptor open with read access.

(`get-lock-region` *fdes lock*) ⟶ *lock-region or* `#f` procedure

Return the first lock region on *fdes* that would conflict with lock region *lock*. If there is no such lock region, return false. This procedure fills out the `proc` field of the returned lock region, and is the only procedure that has anything to do with this field. (See section 3.4.1 for a description of process objects.) Note that if you apply this procedure to a file system that is shared across multiple operating systems (*i.e.*, an NFS file system), the `proc` field may be ambiguous. We note, again, that POSIX advisory file locking is not a terribly useful or well-designed facility.

(`unlock-region` *fdes lock*) ⟶ *undefined* procedure

Release a lock from a file.

(`with-region-lock*` *fdes lock thunk*) ⟶ *value(s) of thunk* procedure
(`with-region-lock` *fdes lock body ...*) ⟶ *value(s) of body* syntax

This procedure obtains the requested lock, and then calls (*thunk*). When *thunk* returns, the lock is released. A non-local exit (*e.g.*, throwing to a saved continuation or raising an exception) also causes the lock to be released.

After a normal return from *thunk*, its return values are returned by `with-region-lock*`. The `with-region-lock` special form is equivalent syntactic sugar.

3.3 File system

Besides the following procedures, which allow access to the computer's file system, scsh also provides a set of procedures which manipulate file *names*. These string-processing procedures are documented in section 5.1.1.

(`create-directory` *fname [perms override?]*) ⟶ *undefined* procedure
(`create-fifo` *fname [perms override?]*) ⟶ *undefined* procedure
(`create-hard-link` *oldname newname [override?]*) ⟶ *undefined* procedure
(`create-symlink` *old-name new-name [override?]*) ⟶ *undefined* procedure

These procedures create objects of various kinds in the file system.

The *override?* argument controls the action if there is already an object in the file system with the new name:

`#f`	signal an error (default)
`'query`	prompt the user
other	delete the old object (with `delete-file` or `delete-directory`, as appropriate) before creating the new object.

Perms defaults to `#o777` (but is masked by the current umask).

> *Remark:* Currently, if you try to create a hard or symbolic link from a file to itself, you will error out with *override?* false, and simply delete your file with *override?* true. Catching this will require some sort of true-name procedure, which I currently do not have.

(`delete-directory` *fname*) ⟶ *undefined* procedure
(`delete-file` *fname*) ⟶ *undefined* procedure
(`delete-filesys-object` *fname*) ⟶ *undefined* procedure

These procedures delete objects from the file system. The `delete-filesys-object` procedure will delete an object of any type from the file system: files, (empty) directories, symlinks, fifos, *etc.*.

If the object being deleted doesn't exist, `delete-directory` and `delete-file` raise an error, while `delete-filesys-object` simply returns.

(`read-symlink` *fname*) ⟶ *string* procedure

Return the filename referenced by symbolic link `fname`.

(`rename-file` *old-fname new-fname [override?]*) ⟶ *undefined* procedure

If you override an existing object, then *old-fname* and *new-fname* must type-match—either both directories, or both non-directories. This is required by the semantics of Unix `rename()`.

> *Remark:* There is an unfortunate atomicity problem with the `rename-file` procedure: if you specify no-override, but create file `new-fname` sometime between `rename-file`'s existence check and the actual rename operation, your file will be clobbered with `old-fname`. There is no way to fix this problem, given the semantics of Unix `rename()`; at least it is highly unlikely to occur in practice.

(`set-file-mode` *fname/fd/port mode*) ⟶ *undefined* procedure
(`set-file-owner` *fname/fd/port uid*) ⟶ *undefined* procedure
(`set-file-group` *fname/fd/port gid*) ⟶ *undefined* procedure

These procedures set the permission bits, owner id, and group id of a file, respectively. The file can be specified by giving the file name, or either an integer file descriptor or a port open on the file. Setting file user ownership usually requires root privileges.

(`set-file-times` *fname [access-time mod-time]*) ⟶ *undefined* procedure

This procedure sets the access and modified times for the file *fname* to the supplied values (see section 3.10 for the scsh representation of time). If neither time argument is supplied, they are both taken to be the current time. You must provide both times or neither. If the procedure completes successfully, the file's time of last status-change (`ctime`) is set to the current time.

(`sync-file` *fd/port*) ⟶ *undefined* procedure
(`sync-file-system`) ⟶ *undefined* procedure

Calling `sync-file` causes Unix to update the disk data structures for a given file. If *fd/port* is a port, any buffered data it may have is first flushed. Calling `sync-file-system` synchronises the kernel's entire file system with the disk.

These procedures are not POSIX. Interestingly enough, `sync-file-system` doesn't actually do what it is claimed to do. We just threw it in for humor value. See the `sync(2)` man page for Unix enlightenment.

(`truncate-file` *fname/fd/port len*) ⟶ *undefined* procedure

The specified file is truncated to *len* bytes in length.

(`file-info` *fname/fd/port [chase?]*) ⟶ *file-info-record* procedure

The `file-info` procedure returns a record structure containing everything there is to know about a file. If the *chase?* flag is true (the default), then the procedure chases symlinks and reports on the files to which they refer. If *chase?* is false, then the procedure checks the actual file itself, even if it's a symlink. The *chase?* flag is ignored if the file argument is a file descriptor or port.

The value returned is a *file-info record*, defined to have the following structure:

```
(define-record file-info
  type      ; {block-special, char-special, directory,
            ;     fifo, regular, socket, symlink}
  device    ; Device file resides on.
  inode     ; File's inode.
  mode      ; File's mode bits: permissions, setuid, setgid
  nlinks    ; Number of hard links to this file.
  uid       ; Owner of file.
  gid       ; File's group id.
  size      ; Size of file, in bytes.
  atime     ; Time of last access.
  mtime     ; Time of last mod.
  ctime)    ; Time of last status change.
```

The uid field of a file-info record is accessed with the procedure

```
(file-info:uid x)
```

and similarly for the other fields. The `type` field is a symbol; all other fields are integers. A file-info record is discriminated with the `file-info?` predicate.

The following procedures all return selected information about a file; they are built on top of `file-info`, and are called with the same arguments that are passed to it.

Procedure	returns
`file-type`	type
`file-inode`	inode
`file-mode`	mode
`file-nlinks`	nlinks
`file-owner`	uid
`file-group`	gid
`file-size`	size
`file-last-access`	atime
`file-last-mod`	mtime
`file-last-status-change`	ctime

Example:

```
;; All my files in /usr/tmp:
(filter (λ (f) (= (file-owner f) (user-uid)))
        (directory-files "/usr/tmp")))
```

> *Remark:* `file-info` was named `file-attributes` in releases of scsh prior to release 0.4. We changed the name to `file-info` for consistency with the other information-retrieval procedures in scsh: `user-info`, `group-info`, `host-info`, `network-info` , `service-info`, and `protocol-info`.
>
> The `file-attributes` binding is still supported in the current release of scsh, but is deprecated, and may go away in a future release.

(`file-directory?` *fname/fd/port [chase?]*) ⟶ *boolean* — procedure
(`file-fifo?` *fname/fd/port [chase?]*) ⟶ *boolean* — procedure
(`file-regular?` *fname/fd/port [chase?]*) ⟶ *boolean* — procedure
(`file-socket?` *fname/fd/port [chase?]*) ⟶ *boolean* — procedure
(`file-special?` *fname/fd/port [chase?]*) ⟶ *boolean* — procedure
(`file-symlink?` *fname/fd/port*) ⟶ *boolean* — procedure

These procedures are file-type predicates that test the type of a given file. They are applied to the same arguments to which `file-info` is applied; the sole exception is `file-symlink?`, which does not take the optional *chase?* second argument.

For example,

```
(file-directory? "/usr/dalbertz")   ⟹   #t
```

(`file-not-readable?` *fname*) ⟶ *boolean* — procedure
(`file-not-writable?` *fname*) ⟶ *boolean* — procedure
(`file-not-executable?` *fname*) ⟶ *boolean* — procedure

Returns:

Value	meaning
`#f`	Access permitted
`'search-denied`	Can't stat—a protected directory is blocking access.
`'permission`	Permission denied.
`'no-directory`	Some directory doesn't exist.
`'nonexistent`	File doesn't exist.

A file is considered writeable if either (1) it exists and is writeable or (2) it doesn't exist and the directory is writeable. Since symlink permission bits are ignored by the filesystem, these calls do not take a *chase?* flag.

Note that these procedures use the process' *effective* user and group ids for permission checking. POSIX defines an `access()` function that uses the process' real uid and gids. This is handy for setuid programs that would like to find out if the actual user has specific rights; scsh ought to provide this functionality (but doesn't at the current time).

There are several problems with these procedures. First, there's an atomicity issue. In between checking permissions for a file and then trying an operation on the file, another process could change the permissions, so a return value from these functions guarantees nothing. Second, the code special-cases permission checking when the uid is root—if the file exists, root is assumed to have the requested permission. However, not even root can write a file that is on a read-only file system, such as a CD ROM. In this case, `file-not-writable?` will lie, saying that root has write access, when in fact the opening the file for write access will fail. Finally, write permission confounds write access and create access. These should be disentangled.

Some of these problems could be avoided if POSIX had a real-uid variant of the `access()` call we could use, but the atomicity issue is still a problem. In the final analysis, the only way to find out if you have the right to perform an operation on a file is to try and open it for the desired operation. These permission-checking functions are mostly intended for script-writing, where loose guarantees are tolerated.

(`file-readable?` *fname*) $\longrightarrow$ *boolean* procedure
(`file-writable?` *fname*) $\longrightarrow$ *boolean* procedure
(`file-executable?` *fname*) $\longrightarrow$ *boolean* procedure

These procedures are the logical negation of the preceding `file-not-...?` procedures. Refer to them for a discussion of their problems and limitations.

(`file-not-exists?` *fname [chase?]*) $\longrightarrow$ *object* procedure

Returns:

`#f`	Exists.
`#t`	Doesn't exist.
`'search-denied`	Some protected directory is blocking the search.

(`file-exists?` *fname [chase?]*) $\longrightarrow$ *boolean* procedure

This is simply `(not (file-not-exists?` *fname [chase?]*`))`

`(directory-files` *[dir dotfiles?]*`)` $\longrightarrow$ *string list* procedure

Return the list of files in directory *dir*, which defaults to the current working directory. The *dotfiles?* flag (default `#f`) causes dot files to be included in the list. Regardless of the value of *dotfiles?*, the two files `.` and `..` are *never* returned.

The directory *dir* is not prepended to each file name in the result list. That is,

```
(directory-files "/etc")
```

returns

```
("chown" "exports" "fstab" ...)
```

not

```
("/etc/chown" "/etc/exports" "/etc/fstab" ...)
```

To use the files in returned list, the programmer can either manually prepend the directory:

```
(map (λ (f) (string-append dir "/" f)) files)
```

or cd to the directory before using the file names:

```
(with-cwd dir
  (for-each delete-file (directory-files)))
```

or use the `glob` procedure, defined below.

A directory list can be generated by `(run/strings (ls))`, but this is unreliable, as filenames with whitespace in their names will be split into separate entries. Using `directory-files` is reliable.

`(glob` pat_1 ...`)` $\longrightarrow$ *string list* procedure

Glob each pattern against the filesystem and return the sorted list. Duplicates are not removed. Patterns matching nothing are not included literally.[2] C shell `{a,b,c}` patterns are expanded. Backslash quotes characters, turning off the special meaning of `{`, `}`, `*`, `[`, `]`, and `?`.

Note that the rules of backslash for Scheme strings and glob patterns work together to require four backslashes in a row to specify a single literal backslash. Fortunately, this should be a rare occurrence.

A glob subpattern will not match against dot files unless the first character of the subpattern is a literal "`.`". Further, a dot subpattern will

[2] Why bother to mention such a silly possibility? Because that is what sh does.

not match the files `.` or `..` unless it is a constant pattern, as in `(glob "../*/*.c")`. So a directory's dot files can be reliably generated with the simple glob pattern `".*"`.

Some examples:

```
(glob "*.c" "*.h")
    ;; All the C and #include files in my directory.

(glob "*.c" "*/*.c")
    ;; All the C files in this directory and
    ;; its immediate subdirectories.

(glob "lexer/*.c" "parser/*.c")
(glob "{lexer,parser}/*.c")
    ;; All the C files in the lexer and parser dirs.

(glob "\\{lexer,parser\\}/*.c")
    ;; All the C files in the strange
    ;; directory "{lexer,parser}".

(glob "*\\*")
    ;; All the files ending in "*", e.g.
    ;; ("foo*" "bar*")

(glob "*lexer*")
    ("mylexer.c" "lexer1.notes")
    ;; All files containing the string "lexer".

(glob "lexer")
    ;; Either ("lexer") or ().
```

If the first character of the pattern (after expanding braces) is a slash, the search begins at root; otherwise, the search begins in the current working directory.

If the last character of the pattern (after expanding braces) is a slash, then the result matches must be directories, *e.g.*,

```
(glob "/usr/man/man?/") ⟹
        ("/usr/man/man1/" "/usr/man/man2/" ...)
```

Globbing can sometimes be useful when we need a list of a directory's files where each element in the list includes the pathname for the file. Compare:

```
(directory-files "../include") ⟹
    ("cig.h" "decls.h" ...)

(glob "../include/*") ⟹
    ("../include/cig.h" "../include/decls.h" ...)
```

(glob-quote *str*) ⟶ *string* procedure

Returns a constant glob pattern that exactly matches *str*. All wild-card characters in *str* are quoted with a backslash.

```
(glob-quote "Any *.c files?")
    ⟹ "Any \*.c files\?"
```

(file-match *root dot-files?* pat_1 pat_2 … pat_n) ⟶ *string list* procedure

{Note This procedure is deprecated, and will probably either go away or be substantially altered in a future release. New code should not call this procedure. The problem is that it relies upon Posix-notation regular expressions; the rest of scsh has been converted over to the new SRE notation.}

`file-match` provides a more powerful file-matching service, at the expense of a less convenient notation. It is intermediate in power between most shell matching machinery and recursive `find(1)`.

Each pattern is a regexp. The procedure searches from *root*, matching the first-level files against pattern pat_1, the second-level files against pat_2, and so forth. The list of files matching the whole path pattern is returned, in sorted order. The matcher uses Spencer's regular expression package.

The files `.` and `..` are never matched. Other dot files are only matched if the *dot-files?* argument is `#t`.

A given pat_i pattern is matched as a regexp, so it is not forced to match the entire file name. *E.g.*, pattern `"t"` matches any file containing a "t" in its name, while pattern `"^t$"` matches only a file whose entire name is "`t`".

The pat_i patterns can be more general than stated above.

- A single pattern can specify multiple levels of the path by embedding `/` characters within the pattern. For example, the pattern `"a/b/c"` gives a match equivalent to the list of patterns `"a"` `"b"` `"c"`.

- A pat_i pattern can be a procedure, which is used as a match predicate. It will be repeatedly called with a candidate file-name to test. The file-name will be the entire path accumulated. If the procedure raises an error condition, `file-match` will catch the error and treat it as a failed match. This keeps `file-match` from being blown out of the water by applying tests to dangling symlinks and other similar situations.

Some examples:

```
(file-match "/usr/lib" #f "m$" "^tab") ⟹
    ("/usr/lib/term/tab300" "/usr/lib/term/tab300-12" ...)

(file-match "." #f  "^lex|parse|codegen$" "\\.c$") ⟹
    ("lex/lex.c" "lex/lexinit.c" "lex/test.c"
     "parse/actions.c" "parse/error.c" parse/test.c"
     "codegen/io.c" "codegen/walk.c")

(file-match "." #f  "^lex|parse|codegen$/\\.c$")
     ;; The same.

(file-match "." #f  file-directory?)
    ;; Return all subdirs of the current directory.

(file-match "/" #f  file-directory?) ⟹
    ("/bin" "/dev" "/etc" "/tmp" "/usr")
    ;; All subdirs of root.

(file-match "." #f  "\\.c")
    ;; All the C files in my directory.

(define (ext extension)
  (λ (fn) (string-suffix? fn extension)))

(define (true . x) #t)

(file-match "." #f  "./\\.c")
(file-match "." #f  "" "\\.c")
(file-match "." #f  true "\\.c")
(file-match "." #f  true (ext "c"))
    ;; All the C files of all my immediate subdirs.

(file-match "." #f "lexer") ⟹
    ("mylexer.c" "lexer.notes")
    ;; Compare with (glob "lexer"), above.
```

Note that when *root* is the current working directory ("."), when it is converted to directory form, it becomes "", and doesn't show up in the result file-names.

It is regrettable that the regexp wild card char, "`.`", is such an important file name literal, as dot-file prefix and extension delimiter.

(`create-temp-file` *[prefix]*) ⟶ *string* procedure

`Create-temp-file` creates a new temporary file and return its name. The optional argument specifies the filename prefix to use, and defaults to `"/usr/tmp/`*pid*`"`, where *pid* is the current process' id. The procedure generates a sequence of filenames that have *prefix* as a common prefix, looking for a filename that doesn't already exist in the file system. When it finds one, it creates it, with permission `#o600` and returns the filename. (The file permission can be changed to a more permissive permission with `set-file-mode` after being created).

This file is guaranteed to be brand new. No other process will have it open. This procedure does not simply return a filename that is very likely to be unused. It returns a filename that definitely did not exist at the moment `create-temp-file` created it.

It is not necessary for the process' pid to be a part of the filename for the uniqueness guarantees to hold. The pid component of the default prefix simply serves to scatter the name searches into sparse regions, so that collisions are less likely to occur. This speeds things up, but does not affect correctness.

Security note: doing `i/o` to files created this way in `/usr/tmp/` is not necessarily secure. General users have write access to `/usr/tmp/`, so even if an attacker cannot access the new temp file, he can delete it and replace it with one of his own. A subsequent open of this filename will then give you his file, to which he has access rights. There are several ways to defeat this attack,

1. Use `temp-file-iterate`, below, to return the file descriptor allocated when the file is opened. This will work if the file only needs to be opened once.
2. If the file needs to be opened twice or more, create it in a protected directory, *e.g.*, `$HOME`.
3. Ensure that `/usr/tmp` has its sticky bit set. This requires system administrator privileges.

The actual default prefix used is controlled by the dynamic variable `*temp-file-template*`, and can be overridden for increased security. See `temp-file-iterate`.

`(temp-file-iterate` *maker [template]*`)` $\longrightarrow$ *object*$^+$ procedure
`*temp-file-template*` *string*

This procedure can be used to perform certain atomic transactions on the file system involving filenames. Some examples:

- Linking a file to a fresh backup temp name.
- Creating and opening an unused, secure temp file.
- Creating an unused temporary directory.

This procedure uses *template* to generate a series of trial file names. *Template* is a `format` control string, and defaults to

`"/usr/tmp/`*pid*`.~a"`

where *pid* is the current process' process id. File names are generated by calling `format` to instantiate the template's `~a` field with a varying string.

Maker is a procedure which is serially called on each file name generated. It must return at least one value; it may return multiple values. If the first return value is `#f` or if *maker* raises the `errno/exist` errno exception, `temp-file-iterate` will loop, generating a new file name and calling *maker* again. If the first return value is true, the loop is terminated, returning whatever value(s) *maker* returned.

After a number of unsuccessful trials, `temp-file-iterate` may give up and signal an error.

Thus, if we ignore its optional *prefix* argument, `create-temp-file` could be defined as:

```
(define (create-temp-file)
  (let ((flags (bitwise-ior open/create open/exclusive)))
    (temp-file-iterate
        (λ (f)
          (close (open-output-file f flags #o600))
          f))))
```

To rename a file to a temporary name:

```
(temp-file-iterate (λ (backup)
                     (create-hard-link old-file backup)
                     backup)
                   ".#temp.~a") ; Keep link in cwd.
(delete-file old-file)
```

Recall that scsh reports syscall failure by raising an error exception, not by returning an error code. This is critical to to this example—the programmer can assume that if the `temp-file-iterate` call returns, it returns successully. So the following `delete-file` call can be reliably invoked, safe in the knowledge that the backup link has definitely been established.

To create a unique temporary directory:

```
(temp-file-iterate (λ (dir) (create-directory dir) dir)
                   "/usr/tmp/tempdir.~a")
```

Similar operations can be used to generate unique symlinks and fifos, or to return values other than the new filename (*e.g.*, an open file descriptor or port).

The default template is in fact taken from the value of the dynamic variable `*temp-file-template*`, which itself defaults to `"/usr/tmp/`*pid*`.~a"`, where *pid* is the scsh process' pid. For increased security, a user may wish to change the template to use a directory not allowing world write access (*e.g.*, his home directory).

`(temp-file-channel)` ⟶ *[inp outp]* procedure

This procedure can be used to provide an interprocess communications channel with arbitrary-sized buffering. It returns two values, an input port and an output port, both open on a new temp file. The temp file itself is deleted from the Unix file tree before `temp-file-channel` returns, so the file is essentially unnamed, and its disk storage is reclaimed as soon as the two ports are closed.

`Temp-file-channel` is analogous to `port-pipe` with two exceptions:

- If the writer process gets ahead of the reader process, it will not hang waiting for some small pipe buffer to drain. It will simply buffer the data on disk. This is good.
- If the reader process gets ahead of the writer process, it will also not hang waiting for data from the writer process. It will simply see and report an end of file. This is bad.

  In order to ensure that an end-of-file returned to the reader is legitimate, the reader and writer must serialise their i/o. The simplest way to do this is for the reader to delay doing input until the writer has completely finished doing output, or exited.

3.4 Processes

(`exec` *prog arg1 ... argn*) $\longrightarrow$ *no return value* — procedure
(`exec-path` *prog arg1 ... argn*) $\longrightarrow$ *no return value* — procedure
(`exec/env` *prog env arg1 ... argn*) $\longrightarrow$ *no return value* — procedure
(`exec-path/env` *prog env arg1 ... argn*) $\longrightarrow$ *no return value* — procedure

The `.../env` variants take an environment specified as a string→string alist. An environment of `#t` is taken to mean the current process' environment (*i.e.*, the value of the external char `**environ`).

[Rationale: `#f` is a more convenient marker for the current environment than `#t`, but would cause an ambiguity on Schemes that identify `#f` and `()`.]

The path-searching variants search the directories in the list `exec-path-list` for the program. A path-search is not performed if the program name contains a slash character—it is used directly. So a program with a name like `"bin/prog"` always executes the program `bin/prog` in the current working directory. See `$path` and `exec-path-list`, below.

Note that there is no analog to the C function `execv()`. To get the effect just do

```
(apply exec prog arglist)
```

All of these procedures flush buffered output and close unrevealed ports before executing the new binary. To avoid flushing buffered output, see `%exec` below.

Note that the C `exec()` procedure allows the zeroth element of the argument vector to be different from the file being executed, *e.g.*

```
char *argv[] = {"-", "-f", 0};
exec("/bin/csh", argv, envp);
```

The scsh `exec`, `exec-path`, `exec/env`, and `exec-path/env` procedures do not give this functionality—element 0 of the arg vector is always identical to the `prog` argument. In the rare case the user wishes to differentiate these two items, he can use the low-level `%exec` and `exec-path-search` procedures. These procedures never return under any circumstances. As with any other system call, if there is an error, they raise an exception.

(`%exec` *prog arglist env*) $\longrightarrow$ *undefined* — procedure
(`exec-path-search` *fname pathlist*) $\longrightarrow$ *string or* `#f` — procedure

The `%exec` procedure is the low-level interface to the system call. The *arglist* parameter is a list of arguments; *env* is either a string→string alist or `#t`. The new program's `argv[0]` will be taken from (`car` *arglist*), *not* from *prog*. An environment of `#t` means the current process' environment. `%exec` does not flush buffered output (see `flush-all-ports`).

All exec procedures, including `%exec`, coerce the `prog` and `arg` values to strings using the usual conversion rules: numbers are converted to decimal numerals, and symbols converted to their print-names.

`exec-path-search` searches the directories of *pathlist* looking for an occurrence of file `fname`. If no executable file is found, it returns `#f`. If `fname` contains a slash character, the path search is short-circuited, but the procedure still checks to ensure that the file exists and is executable—if not, it still returns `#f`. Users of this procedure should be aware that it invites a potential race condition: between checking the file with `exec-path-search` and executing it with `%exec`, the file's status might change. The only atomic way to do the search is to loop over the candidate file names, exec'ing each one and looping when the exec operation fails.

See `$path` and `exec-path-list`, below.

`(exit` *[status]*`)` $\longrightarrow$ *no return value* — procedure
`(%exit` *[status]*`)` $\longrightarrow$ *no return value* — procedure

These procedures terminate the current process with a given exit status. The default exit status is 0. The low-level `%exit` procedure immediately terminates the process without flushing buffered output.

`(call-terminally` *thunk*`)` $\longrightarrow$ *no return value* — procedure

`call-terminally` calls its thunk. When the thunk returns, the process exits. Although `call-terminally` could be implemented as

```
(λ (thunk) (thunk) (exit 0))
```

an implementation can take advantage of the fact that this procedure never returns. For example, the runtime can start with a fresh stack and also start with a fresh dynamic environment, where shadowed bindings are discarded. This can allow the old stack and dynamic environment to be collected (assuming this data is not reachable through some live continuation).

`(suspend)` $\longrightarrow$ *undefined* — procedure

Suspend the current process with a SIGSTOP signal.

`(fork` *[thunk]*`)` $\longrightarrow$ *proc or* `#f` — procedure
`(%fork` *[thunk]*`)` $\longrightarrow$ *proc or* `#f` — procedure

`fork` with no arguments is like C `fork()`. In the parent process, it returns the child's *process object* (see below for more information on process objects). In the child process, it returns `#f`.

`fork` with an argument only returns in the parent process, returning the child's process object. The child process calls *thunk* and then exits.

`fork` flushes buffered output before forking, and sets the child process to non-interactive. `%fork` does not perform this bookkeeping; it simply forks.

(`fork/pipe` *[thunk]*) ⟶ *proc or* `#f` procedure
(`%fork/pipe` *[thunk]*) ⟶ *proc or* `#f` procedure

Like `fork` and `%fork`, but the parent and child communicate via a pipe connecting the parent's stdin to the child's stdout. These procedures side-effect the parent by changing his stdin.

In effect, `fork/pipe` splices a process into the data stream immediately upstream of the current process. This is the basic function for creating pipelines. Long pipelines are built by performing a sequence of `fork/pipe` calls. For example, to create a background two-process pipe `a | b`, we write:

```
(fork (λ () (fork/pipe a) (b)))
```

which returns the process object for `b`'s process.

To create a background three-process pipe `a | b | c`, we write:

```
(fork (λ () (fork/pipe a)
            (fork/pipe b)
            (c)))
```

which returns the process object for `c`'s process.

Note that these procedures affect file descriptors, not ports. That is, the pipe is allocated connecting the child's file descriptor 1 to the parent's file descriptor 0. *Any previous Scheme port built over these affected file descriptors is shifted to a new, unused file descriptor with* `dup` *before allocating the I/O pipe.* This means, for example, that the ports bound to `(current-input-port)` and `(current-output-port)` in either process are not affected—they still refer to the same I/O sources and sinks as before. Remember the simple scsh rule: Scheme ports are bound to I/O sources and sinks, *not* particular file descriptors.

If the child process wishes to rebind the current output port to the pipe on file descriptor 1, it can do this using `with-current-output-port` or a related form. Similarly, if the parent wishes to change the current input port to the pipe on file descriptor 0, it can do this using set-current-input-port! or a related form. Here is an example showing how to set up the I/O ports on both sides of the pipe:

```
(fork/pipe (λ ()
             (with-current-output-port (fdes->outport 1)
               (display "Hello, world.\n"))))

(set-current-input-port! (fdes->inport 0)
(read-line)          ; Read the string output by the child.
```

None of this is necessary when the I/O is performed by an exec'd program in the child or parent process, only when the pipe will be referenced by Scheme code through one of the default current I/O ports.

(fork/pipe+ *conns [thunk]*) ⟶ *proc or* #f — procedure
(%fork/pipe+ *conns [thunk]*) ⟶ *proc or* #f — procedure

Like fork/pipe, but the pipe connections between the child and parent are specified by the connection list *conns*. See the

(|+ *conns* pf_1 ... pf_n)

process form for a description of connection lists.

3.4.1 Process objects and process reaping

Scsh uses *process objects* to represent Unix processes. They are created by the fork procedure, and have the following exposed structure:

```
(define-record proc
       pid)
```

The only exposed slot in a proc record is the process' pid, the integer id assigned by Unix to the process. The only exported primitive procedures for manipulating process objects are proc? and proc:pid. Process objects are created with the fork procedure.

(pid->proc *pid [probe?]*) ⟶ *proc* — procedure

This procedure maps integer Unix process ids to scsh process objects. It is intended for use in interactive and debugging code, and is deprecated for use in production code. If there is no process object in the system indexed by the given pid, pid->proc's action is determined by the *probe?* parameter (default #f):

probe?	Return
#f	*signal error condition.*
'create	Create new proc object.
True value	#f

Sometime after a child process terminates, scsh will perform a `wait` system call on the child in background, caching the process' exit status in the child's proc object. This is called "reaping" the process. Once the child has been waited, the Unix kernel can free the storage allocated for the dead process' exit information, so process reaping prevents the process table from becoming cluttered with un-waited dead child processes (a.k.a. "zombies"). This can be especially severe if the scsh process never waits on child processes at all; if the process table overflows with forgotten zombies, the OS may be unable to fork further processes.

Reaping a child process moves its exit status information from the kernel into the scsh process, where it is cached inside the child's process object. If the scsh user drops all pointers to the process object, it will simply be garbage collected. On the other hand, if the scsh program retains a pointer to the process object, it can use scsh's `wait` system call to synchronise with the child and retrieve its exit status multiple times (this is not possible with simple Unix integer pids in C—the programmer can only wait on a pid once).

Thus, process objects allow scsh programmer to do two things not allowed in other programming environments:

- Subprocesses that are never waited on are still removed from the process table, and their associated exit status data is eventually automatically garbage collected.
- Subprocesses can be waited on multiple times.

However, note that once a child has exited, if the scsh programmer drops all pointers to the child's proc object, the child's exit status will be reaped and thrown away. This is the intended behaviour, and it means that integer pids are not enough to cause a process's exit status to be retained by the scsh runtime. (This is because it is clearly impossible to GC data referenced by integers.)

As a convenience for interactive use and debugging, all procedures that take process objects will also accept integer Unix pids as arguments, coercing them to the corresponding process objects. Since integer process ids are not reliable ways to keep a child's exit status from being reaped and garbage collected, programmers are encouraged to use process objects in production code.

`(autoreap-policy` *[policy]*`)` $\longrightarrow$ *old-policy* procedure

The scsh programmer can choose different policies for automatic process reaping. The policy is determined by applying this procedure to one of the values `'early`, `'late`, or `#f` (*i.e.*, no autoreap).

early The child is reaped from the Unix kernel's process table into scsh as soon as it dies. This is done by having a signal handler for the `SIGCHLD` signal reap the process. *If a scsh program sets its own handler for the* `SIGCHLD` *signal, the handler must reap dead children by calling* `wait`*,* `wait-any`*, or* `reap-zombies`*.* We deprecate interrupt-driven code, and hope to provide alternative tools in a future, multi-threaded release of scsh.

late The child is not autoreaped until it dies *and* the scsh program drops all pointers to its process object. That is, the process table is cleaned out during garbage collection.

> *Oops:* The `late` policy is not supported under the current release of scsh. It requires more sophisticated gc hooks than we can get from the release of Scheme 48 that we use.

`#f` If autoreaping is turned off, process reaping is completely under control of the programmer, who can force outstanding zombies to be reaped by manually calling the `reap-zombies` procedure (see below).

Note that under any of the autoreap policies, a particular process p can be manually reaped into scsh by simply calling `(wait` p`)`. *All* zombies can be manually reaped with `reap-zombies`.

The `autoreap-policy` procedure returns the policy's previous value. Calling `autoreap-policy` with no arguments returns the current policy without no change.

`(reap-zombies)` $\longrightarrow$ *boolean* procedure

This procedure reaps all outstanding exited child processes into scsh. It returns true if there are no more child processes to wait on, and false if there are outstanding processes still running or suspended.

Issues with process reaping

Reaping a process does not reveal its process group at the time of death; this information is lost when the process reaped. This means that a dead, reaped process is *not eligible* as a return value for a future `wait-process-group` call. This is not likely to be a problem for most code, as programs almost never wait on exited processes by process group. Process group waiting is usually applied to *stopped* processes, which are never reaped. So it is unlikely that this will be a problem for most programs.

Automatic process reaping is a useful programming convenience. However, if a program is careful to wait for all children, and does not wish automatic reaping to happen, the programmer can simply turn process autoreaping off.

Programs that do not wish to use automatic process reaping should be aware that some scsh routines create subprocesses but do not return the child's pid: `run/port*`, and its related procedures and special forms (`run/strings`, *et al.*). Automatic process reaping will clean the child processes created by these procedures out of the kernel's process table. If a program doesn't use process reaping, it should either avoid these forms, or use `wait-any` to wait for the children to exit.

3.4.2 Process waiting

(`wait` *proc/pid [flags]*) $\longrightarrow$ *status* — procedure

This procedure waits until a child process exits, and returns its exit code. The *proc/pid* argument is either a process object (section 3.4.1) or an integer process id. `Wait` returns the child's exit status code (or suspension code, if the `wait/stopped-children` option is used, see below). Status values can be queried with the procedures in section 3.4.3.

The *flags* argument is an integer whose bits specify additional options. It is composed by or'ing together the following flags:

Flag	Meaning
`wait/poll`	Return `#f` immediately if child still active.
`wait/stopped-children`	Wait for suspend as well as exit.

(`wait-any` *[flags]*) $\longrightarrow$ *[proc status]* — procedure

The optional *flags* argument is as for `wait`. This procedure waits for any child process to exit (or stop, if the `wait/stopped-children` flag is used) It returns the process' process object and status code. If there are no children left for which to wait, the two values `[#f #t]` are returned. If the `wait/poll` flag is used, and none of the children are immediately eligble for waiting, then the values `[#f #f]` are returned:

`[#f #f]`	Poll, none ready
`[#f #t]`	No children

`Wait-any` will not return a process that has been previously waited by any other process-wait procedure (`wait`, `wait-any`, and `wait-process-group`). It will return reaped processes that haven't yet been waited.

The use of `wait-any` is deprecated.

(`wait-process-group` *proc/pid [flags]*) $\longrightarrow$ *[proc status]* — procedure

This procedure waits for any child whose process group is *proc/pid* (either a process object or a pid). The *flags* argument is as for `wait`.

Note that if the programmer wishes to wait for exited processes by process group, the program should take care not to use process reaping (section 3.4.1), as this loses process group information. However, most process-group waiting is for stopped processes (to implement job control), so this is rarely an issue, as stopped processes are not subject to reaping.

3.4.3 Analysing process status codes

When a child process dies (or is suspended), its parent can call the `wait` procedure to recover the exit (or suspension) status of the child. The exit status is a small integer that encodes information describing how the child terminated. The bit-level format of the exit status is not defined by POSIX; you must use the following three functions to decode one. However, if a child terminates normally with exit code 0, POSIX does require `wait` to return an exit status that is exactly zero. So `(zero?` *status*`)` is a correct way to test for non-error, normal termination, *e.g.*,

```
(if (zero? (run (rcp scsh.tar.gz lambda.csd.hku.hk:)))
    (delete-file "scsh.tar.gz"))
```

`(status:exit-val` *status*`)`	⟶	*integer or* `#f`	procedure
`(status:stop-sig` *status*`)`	⟶	*integer or* `#f`	procedure
`(status:term-sig` *status*`)`	⟶	*integer or* `#f`	procedure

For a given status value produced by calling `wait`, exactly one of these routines will return a true value.

If the child process exited normally, `status:exit-val` returns the exit code for the child process (*i.e.*, the value the child passed to `exit` or returned from `main`). Otherwise, this function returns false.

If the child process was suspended by a signal, `status:stop-sig` returns the signal that suspended the child. Otherwise, this function returns false.

If the child process terminated abnormally, `status:term-sig` returns the signal that terminated the child. Otherwise, this function returns false.

3.5 Process state

(umask) ⟶ *fixnum* procedure
(set-umask *perms*) ⟶ *undefined* procedure
(with-umask* *perms thunk*) ⟶ *value(s) of thunk* procedure
(with-umask *perms . body*) ⟶ *value(s) of body* syntax

The process' current umask is retrieved with `umask`, and set with `(set-umask perms)`. Calling `with-umask*` changes the umask to *perms* for the duration of the call to *thunk*. If the program throws out of *thunk* by invoking a continuation, the umask is reset to its external value. If the program throws back into *thunk* by calling a stored continuation, the umask is restored to the *perms* value. The special form `with-umask` is equivalent in effect to the procedure `with-umask*`, but does not require the programmer to explicitly wrap a `(λ () ...)` around the body of the code to be executed.

(chdir *[fname]*) ⟶ *undefined* procedure
(cwd) ⟶ *string* procedure
(with-cwd* *fname thunk*) ⟶ *value(s) of thunk* procedure
(with-cwd *fname . body*) ⟶ *value(s) of body* syntax

These forms manipulate the current working directory. The cwd can be changed with `chdir` (although in most cases, `with-cwd` is preferrable). If `chdir` is called with no arguments, it changes the cwd to the user's home directory. The `with-cwd*` procedure calls `thunk` with the cwd temporarily set to *fname*; when *thunk* returns, or is exited in a non-local fashion (*e.g.*, by raising an exception or by invoking a continuation), the cwd is returned to its original value. The special form `with-cwd` is simply syntactic sugar for `with-cwd*`.

(pid) ⟶ *fixnum* procedure
(parent-pid) ⟶ *fixnum* procedure
(process-group) ⟶ *fixnum* procedure
(set-process-group *[proc/pid] pgrp*) ⟶ *undefined* procedure

`(pid)` and `(parent-pid)` retrieve the process id for the current process and its parent. `(process-group)` returns the process group of the current process. A process' process-group can be set with `set-process-group`; the value *proc/pid* specifies the affected process. It may be either a process object or an integer process id, and defaults to the current process.

(set-priority *which who priority*) ⟶ *undefined* procedure
(priority *which who*) ⟶ *fixnum* procedure
(nice *[proc/pid delta]*) ⟶ *undefined* procedure

These procedures set and access the priority of processes. I can't remember how `set-priority` and `priority` work, so no documentation, and besides, they aren't implemented yet, anyway.

`(user-login-name)` ⟶ *string* procedure
`(user-uid)` ⟶ *fixnum* procedure
`(user-effective-uid)` ⟶ *fixnum* procedure
`(user-gid)` ⟶ *fixnum* procedure
`(user-effective-gid)` ⟶ *fixnum* procedure
`(user-supplementary-gids)` ⟶ *fixnum list* procedure
`(set-uid` *uid*`)` ⟶ *undefined* procedure
`(set-gid` *gid*`)` ⟶ *undefined* procedure

These routines get and set the effective and real user and group ids. The `set-uid` and `set-gid` routines correspond to the POSIX `setuid()` and `setgid()` procedures.

`(process-times)` ⟶ *[fixnum fixnum fixnum fixnum]* procedure

Returns four values:
user CPU time in clock-ticks
system CPU time in clock-ticks
user CPU time of all descendant processes
system CPU time of all descendant processes
Note that CPU time clock resolution is not the same as the real-time clock resolution provided by `time+ticks`. That's Unix.

`(cpu-ticks/sec)` ⟶ *integer* procedure

Returns the resolution of the CPU timer in clock ticks per second. This can be used to convert the times reported by `process-times` to seconds.

3.6 User and group database access

These procedures are used to access the user and group databases (*e.g.*, the ones traditionally stored in `/etc/passwd` and `/etc/group`.)

`(user-info` *uid/name*`)` ⟶ *record* procedure

Return a `user-info` record giving the recorded information for a particular user:

```
(define-record user-info
  name uid gid home-dir shell)
```

The *uid/name* argument is either an integer uid or a string user-name.

(`->uid` *uid/name*) ⟶ *fixnum* procedure
(`->username` *uid/name*) ⟶ *string* procedure

These two procedures coerce integer uid's and user names to a particular form.

(`group-info` *gid/name*) ⟶ *record* procedure

Return a `group-info` record giving the recorded information for a particular group:

```
(define-record group-info
  name gid members)
```

The *gid/name* argument is either an integer gid or a string group-name.

3.7 Accessing command-line arguments

`command-line-arguments` *string* list
(`command-line`) ⟶ *string list* procedure

The list of strings `command-line-arguments` contains the arguments passed to the scsh process on the command line. Calling `(command-line)` returns the complete `argv` string list, including the program. So if we run a scsh program

```
/usr/shivers/bin/myls -CF src
```

then `command-line-arguments` is

```
("-CF" "src")
```

and `(command-line)` returns

```
("/usr/shivers/bin/myls" "-CF" "src")
```

`command-line` returns a fresh list each time it is called. In this way, the programmer can get a fresh copy of the original argument list if `command-line-arguments` has been modified or is lexically shadowed.

(`arg` *arglist n [default]*) ⟶ *string* procedure
(`arg*` *arglist n [default-thunk]*) ⟶ *string* procedure
(`argv` *n [default]*) ⟶ *string* procedure

These procedures are useful for accessing arguments from argument lists. `arg` returns the n^{th} element of *arglist*. The index is 1-based. If *n* is too large, *default* is returned; if no *default*, then an error is signaled.

`arg*` is similar, except that the *default-thunk* is called to generate the default value.

(argv n) is simply (arg (command-line) (+ n 1)). The +1 offset ensures that the two forms

```
(arg command-line-arguments n)
(argv n)
```

return the same argument (assuming the user has not rebound or modified `command-line-arguments`).

Example:

```
(if (null? command-line-arguments)
    (& (xterm -n ,host -title ,host
              -name ,(string-append "xterm_" host)))
    (let* ((progname (file-name-nondirectory (argv 1)))
           (title (string-append host ":" progname)))
      (& (xterm -n     ,title
                -title ,title
                -e     ,@command-line-arguments))))
```

A subtlety: when the scsh interpreter is used to execute a scsh program, the program name reported in the head of the (command-line) list is the scsh program, *not* the interpreter. For example, if we have a shell script in file `fullecho`:

```
#!/usr/local/bin/scsh -s
!#
(for-each (λ (arg) (display arg) (display " "))
          (command-line))
```

and we run the program

```
fullecho hello world
```

the program will print out

```
fullecho hello world
```

not

```
/usr/local/bin/scsh -s fullecho hello world
```

This argument line processing ensures that if a scsh program is subsequently compiled into a standalone executable or byte-compiled to a heap-image executable by the Scheme 48 virtual machine, its semantics will be unchanged—the arglist processing is invariant. In effect, the

```
/usr/local/bin/scsh -s
```

is not part of the program; it's a specification for the machine to execute the program on, so it is not properly part of the program's argument list.

3.8 System parameters

(system-name) ⟶ *string* procedure

Returns the name of the host on which we are executing. This may be a local name, such as "solar," as opposed to a fully-qualified domain name such as "solar.csie.ntu.edu.tw."

3.9 Signal system

Signal numbers are bound to the variables `signal/hup`, `signal/int`, ... See tables 3.9 and 3.3 for the full list.

(signal-process *proc sig*) ⟶ *undefined* procedure
(signal-process-group *prgrp sig*) ⟶ *undefined* procedure

These two procedures send signals to a specific process, and all the processes in a specific process group, respectively. The *proc* and *prgrp* arguments are either processes or integer process ids.

(itimer *???*) ⟶ *undefined* procedure
(pause-until-interrupt) ⟶ *undefined* procedure

(sleep *secs*) ⟶ *undefined* procedure
(sleep-until *time*) ⟶ *undefined* procedure

The `sleep` procedure causes the process to sleep for *secs* seconds. The `sleep-until` procedure causes the process to sleep until *time* (see section 3.10).

Interrupt handlers

Scsh interrupt handlers are complicated by the fact that scsh is implemented on top of the Scheme 48 virtual machine, which has its own interrupt system, independent of the Unix signal system. This means that Unix signals are delivered in two stages: first, Unix delivers the signal to the Scheme 48 virtual machine, then the Scheme 48 virtual machine delivers the signal to the executing Scheme program as a Scheme 48 interrupt. This ensures that signal delivery happens between two vm instructions, keeping individual instructions atomic.

The Scheme 48 machine has its own set of interrupts, which includes the asynchronous Unix signals (table 3.9). Note that scsh does *not* support signal handlers for "synchronous" Unix signals, such as `signal/ill` or `signal/pipe` (see table 3.3). Synchronous occurrences of these signals are better handled by raising a Scheme exception. We recommend you avoid using signal handlers

Interrupt	Unix signal	OS Variant
`interrupt/alrm`[a]	`signal/alrm`	POSIX
`interrupt/int`[b]	`signal/int`	POSIX
`interrupt/memory-shortage`	N/A	
`interrupt/chld`	`signal/chld`	POSIX
`interrupt/cont`	`signal/cont`	POSIX
`interrupt/hup`	`signal/hup`	POSIX
`interrupt/quit`	`signal/quit`	POSIX
`interrupt/term`	`signal/term`	POSIX
`interrupt/tstp`	`signal/tstp`	POSIX
`interrupt/usr1`	`signal/usr1`	POSIX
`interrupt/usr2`	`signal/usr2`	POSIX
`interrupt/info`	`signal/info`	BSD only
`interrupt/io`	`signal/io`	BSD + SVR4
`interrupt/poll`	`signal/poll`	SVR4 only
`interrupt/prof`	`signal/prof`	BSD + SVR4
`interrupt/pwr`	`signal/pwr`	SVR4 only
`interrupt/urg`	`signal/urg`	BSD + SVR4
`interrupt/vtalrm`	`signal/vtalrm`	BSD + SVR4
`interrupt/winch`	`signal/winch`	BSD + SVR4
`interrupt/xcpu`	`signal/xcpu`	BSD + SVR4
`interrupt/xfsz`	`signal/xfsz`	BSD + SVR4

Table 3.2: Scheme 48 virtual-machine interrupts and related Unix signals. Only the POSIX signals are guaranteed to be defined; however, your implementation and OS may define other signals and interrupts not listed here.

[a] Also bound to Scheme 48 interrupt `interrupt/alarm`.

[b] Also bound to Scheme 48 interrupt `interrupt/keyboard`.

Unix signal	Type	OS Variant
`signal/stop`	Uncatchable	POSIX
`signal/kill`	Uncatchable	POSIX
`signal/abrt`	Synchronous	POSIX
`signal/fpe`	Synchronous	POSIX
`signal/ill`	Synchronous	POSIX
`signal/pipe`	Synchronous	POSIX
`signal/segv`	Synchronous	POSIX
`signal/ttin`	Synchronous	POSIX
`signal/ttou`	Synchronous	POSIX
`signal/bus`	Synchronous	BSD + SVR4
`signal/emt`	Synchronous	BSD + SVR4
`signal/iot`	Synchronous	BSD + SVR4
`signal/sys`	Synchronous	BSD + SVR4
`signal/trap`	Synchronous	BSD + SVR4

Table 3.3: Uncatchable and synchronous Unix signals. While these signals may be sent with `signal-process` or `signal-process-group`, there are no corresponding scsh interrupt handlers. Only the POSIX signals are guaranteed to be defined; however, your implementation and OS may define other signals not listed here.

unless you absolutely have to; we intend to provide a better, higher-level interface to Unix signals after scsh has been ported to a multi-threaded platform.

`(signal->interrupt` *integer*`)` $\longrightarrow$ *integer* procedure

The programmer maps from Unix signals to Scheme 48 interrupts with the `signal->interrupt` procedure. If the signal does not have a defined Scheme 48 interrupt, an errror is signaled.

`(interrupt-set` $integer_1 \ldots integer_n$`)` $\longrightarrow$ *integer* procedure

This procedure builds interrupt sets from its interrupt arguments. A set is represented as an integer using a two's-complement representation of the bit set.

`(enabled-interrupts)` $\longrightarrow$ *interrupt-set* procedure
`(set-enabled-interrupts` *interrupt-set*`)` $\longrightarrow$ *interrupt-set* procedure

Get and set the value of the enabled-interrupt set. Only interrupts in this set have their handlers called when delivered. When a disabled interrupt is delivered to the Scheme 48 machine, it is held pending until it becomes enabled, at which time its handler is invoked.

Interrupt sets are represented as integer bit sets (constructed with the `interrupt-set` function). The `set-enabled-interrupts` procedure returns the previous value of the enabled-interrupt set.

`(with-enabled-interrupts` *interrupt-set . body*`)` $\longrightarrow$ *value(s) of body* syntax
`(with-enabled-interrupts*` *interrupt-set thunk*`)` $\longrightarrow$ *value(s) of thunk* procedure

Run code with a given set of interrupts enabled. Note that "enabling" an interrupt means enabling delivery from the Scheme 48 vm to the scsh program. Using the Scheme 48 interrupt system is fairly lightweight, and does not involve actually making a system call. Note that enabling an interrupt means that the assigned interrupt handler is allowed to run when the interrupt is delivered. Interrupts not enabled are held pending when delivered.

Interrupt sets are represented as integer bit sets (constructed with the `interrupt-set` function).

`(set-interrupt-handler` *interrupt handler*`)` $\longrightarrow$ *old-handler* procedure

Assigns a handler for a given interrupt, and returns the interrupt's old handler. The *handler* argument is `#f` (ignore), `#t` (default), or a procedure taking an integer argument; the return value follows the same conventions. Note that the *interrupt* argument is an interrupt value, not a signal value. An interrupt is delivered to the Scheme 48 machine by (1) blocking all interrupts, and (2) applying the handler procedure to the set of interrupts that were enabled prior to the interrupt delivery. If the procedure returns normally (*i.e.*, it doesn't throw to a continuation), the set of enabled interrupts will be returned to its previous value. (To restore the enabled-interrupt set before throwing out of an interrupt handler, see `set-enabled-interrupts`)

{Note If you set a handler for the `interrupt/chld` interrupt, you may break scsh's autoreaping process machinery. See the discussion of autoreaping in section 3.4.1.}

(`interrupt-handler` *interrupt*) $\longrightarrow$ *handler* procedure

Return the handler for a given interrupt. Note that the argument is an interrupt value, not a signal value. A handler is either `#f` (ignore), `#t` (default), or a procedure taking an integer argument.

3.10 Time

Scsh's time system is fairly sophisticated, particularly with respect to its careful treatment of time zones. However, casual users shouldn't be intimidated; all of the complexity is optional, and defaulting all the optional arguments reduces the system to a simple interface.

> *Remark:* This time package does not currently work with NeXTSTEP, as NeXTSTEP does not provide a POSIX-compliant time library that will even link.

3.10.1 Terminology

"UTC" and "UCT" stand for "universal coordinated time," which is the official name for what is colloquially referred to as "Greenwich Mean Time."

POSIX allows a single time zone to specify *two* different offsets from UTC: one standard one, and one for "summer time." Summer time is frequently some sort of daylight savings time.

The scsh time package consistently uses this terminology: we never say "gmt" or "dst;" we always say "utc" and "summer time."

3.10.2 Basic data types

We have two types: *time* and *date*.

A *time* specifies an instant in the history of the universe. It is location and time-zone independent.[3] A time is a real value giving the number of elapsed seconds since the Unix "epoch" (Midnight, January 1, 1970 UTC). Time values provide arbitrary time resolution, limited only by the number system of the underlying Scheme system.

A *date* is a name for an instant in time that is specified relative to some location/time-zone in the world, *e.g.*:
Friday October 31, 1994 3:47:21 pm EST.
Dates provide one-second resolution, and are expressed with the following record type:

```
(define-record date     ; A Posix tm struct
  seconds       ; Seconds after the minute [0-59]
  minute        ; Minutes after the hour [0-59]
  hour          ; Hours since midnight [0-23]
  month-day     ; Day of the month [1-31]
  month         ; Months since January [0-11]
  year          ; Years since 1900
  tz-name       ; Time-zone name: #f or a string.
  tz-secs       ; Time-zone offset: #f or an integer.
  summer?       ; Summer (Daylight Savings) time in effect?
  week-day      ; Days since Sunday [0-6]
  year-day)     ; Days since Jan. 1 [0-365]
```

If the `tz-secs` field is given, it specifies the time-zone's offset from UTC in seconds. If it is specified, the `tz-name` and `summer?` fields are ignored when using the date structure to determine a specific instant in time.

If the `tz-name` field is given, it is a time-zone string such as `"EST"` or `"HKT"` understood by the OS. Since POSIX time-zone strings can specify dual standard/summer time-zones (e.g., "EST5EDT" specifies U.S. Eastern Standard/Eastern Daylight Time), the value of the `summer?` field is used to resolve the amiguous boundary cases. For example, on the morning of the Fall daylight savings change-over, 1:00am–2:00am happens twice. Hence the date 1:30 am on this morning can specify two different seconds; the `summer?` flag says which one.

A date with `tz-name` = `tz-secs` = `#f` is a date that is specified in terms of the system's current time zone.

[3]Physics pedants please note: The scsh authors live in a Newtonian universe. We disclaim responsibility for calculations performed in non-ANSI standard light-cones.

There is redundancy in the `date` data structure. For example, the `year-day` field is redundant with the `month-day` and `month` fields. Either of these implies the values of the `week-day` field. The `summer?` and `tz-name` fields are redundant with the `tz-secs` field in terms of specifying an instant in time. This redundancy is provided because consumers of dates may want it broken out in different ways. The scsh procedures that produce date records fill them out completely. However, when date records produced by the programmer are passed to scsh procedures, the redundancy is resolved by ignoring some of the secondary fields. This is described for each procedure below.

(`make-date` *s min h mday mon y [tzn tzs summ? wday yday]*) $\longrightarrow$ *date* procedure

> When making a `date` record, the last five elements of the record are optional, and default to `#f`, `#f`, `#f`, 0, and 0 respectively. This is useful when creating a `date` record to pass as an argument to `time`.

3.10.3 Time zones

Several time procedures take time zones as arguments. When optional, the time zone defaults to local time zone. Otherwise the time zone can be one of:

`#f`	Local time
Integer	Seconds of offset from UTC. For example, New York City is -18000 (-5 hours), San Francisco is -28800 (-8 hours).
String	A POSIX time zone string understood by the OS (*i.e.*., the sort of time zone assigned to the `$TZ` environment variable).

An integer time zone gives the number of seconds you must add to UTC to get time in that zone. It is *not* "seconds west" of UTC—that flips the sign.

To get UTC time, use a time zone of either 0 or `"UCT0"`.

3.10.4 Procedures

(`time+ticks`) $\longrightarrow$ *[secs ticks]* procedure
(`ticks/sec`) $\longrightarrow$ *real* procedure

> The current time, with sub-second resolution. Sub-second resolution is not provided by POSIX, but is available on many systems. The time is returned as elapsed seconds since the Unix epoch, plus a number of sub-second "ticks." The length of a tick may vary from implementation to implementation; it can be determined from `(ticks/sec)`.
>
> The system clock is not required to report time at the full resolution given by `(ticks/sec)`. For example, on BSD, time is reported at 1μs resolution,

so `(ticks/sec)` is 1,000,000. That doesn't mean the system clock has micro-second resolution.

If the OS does not support sub-second resolution, the *ticks* value is always 0, and `(ticks/sec)` returns 1.

> *Remark:* I chose to represent system clock resolution as ticks/sec instead of sec/tick to increase the odds that the value could be represented as an exact integer, increasing efficiency and making it easier for Scheme implementations that don't have sophisticated numeric support to deal with the quantity.
>
> You can convert seconds and ticks to seconds with the expression
>
> `(+` *secs* `(/` *ticks* `(ticks/sec)))`
>
> Given that, why not have the fine-grain time procedure just return a non-integer real for time? Following Common Lisp, I chose to allow the system clock to report sub-second time in its own units to lower the overhead of determining the time. This would be important for a system that wanted to precisely time the duration of some event. Time stamps could be collected with little overhead, deferring the overhead of precisely calculating with them until after collection.
>
> This is all a bit academic for the Scheme 48 implementation, where we determine time with a heavyweight system call, but it's nice to plan for the future.

`(date)` ⟶ *date-record* — procedure
`(date` *[time tz]*`)` ⟶ *date-record* — procedure

Simple `(date)` returns the current date, in the local time zone.

With the optional arguments, `date` converts the time to the date as specified by the time zone *tz*. *Time* defaults to the current time; *tz* defaults to local time, and is as described in the time-zone section.

If the *tz* argument is an integer, the date's `tz-name` field is a POSIX time zone of the form "`UTC+`*hh*`:`*mm*`:`*ss*"; the trailing `:`*mm*`:`*ss* portion is deleted if it is zeroes.

> *Oops:* The Posix facility for converting dates to times, `mktime()`, has a broken design: it indicates an error by returning -1, which is also a legal return value (for date 23:59:59 UCT, 12/31/1969). Scsh resolves the ambiguity in a paranoid fashion: it always reports an error if the underlying Unix facility returns -1. We feel your pain.

`(time)` ⟶ *integer* — procedure
`(time` *[date]*`)` ⟶ *integer* — procedure

Simple `(time)` returns the current time.

With the optional date argument, `time` converts a date to a time. *Date* defaults to the current date.

Note that the input *date* record is overconstrained. `time` ignores *date*'s `week-day` and `year-day` fields. If the date's `tz-secs` field is set, the `tz-name` and `summer?` fields are ignored.

If the `tz-secs` field is `#f`, then the time-zone is taken from the `tz-name` field. A false `tz-name` means the system's current time zone. When calculating with time-zones, the date's `summer?` field is used to resolve ambiguities:

`#f`	Resolve an ambiguous time in favor of non-summer time.
true	Resolve an ambiguous time in favor of summer time.

This is useful in boundary cases during the change-over. For example, in the Fall, when US daylight savings time changes over at 2:00 am, 1:30 am happens twice—it names two instants in time, an hour apart.

Outside of these boundary cases, the `summer?` flag is ignored. For example, if the standard/summer change-overs happen in the Fall and the Spring, then the value of `summer?` is ignored for a January or July date. A January date would be resolved with standard time, and a July date with summer time, regardless of the `summer?` value.

The `summer?` flag is also ignored if the time zone doesn't have a summer time—for example, simple UTC.

`(date->string` *date*`)` $\longrightarrow$ *string* procedure
`(format-date` *fmt date*`)` $\longrightarrow$ *string* procedure

`Date->string` formats the date as a 24-character string of the form:
Sun Sep 16 01:03:52 1973

`Format-date` formats the date according to the format string *fmt*. The format string is copied verbatim, except that tilde characters indicate conversion specifiers that are replaced by fields from the date record. Figure 3.1 gives the full set of conversion specifiers supported by `format-date`.

`(fill-in-date!` *date*`)` $\longrightarrow$ *date* procedure

This procedure fills in missing, redundant slots in a date record. In decreasing order of priority:

- **year, month, month-day ⇒ year-day**
  If the `year`, `month`, and `month-day` fields are all defined (are all integers), the `year-day` field is set to the corresponding value.
- **year, year-day ⇒ month, month-day**
  If the `month` and `month-day` fields aren't set, but the `year` and `year-day` fields are set, then `month` and `month-day` are calculated.

`~~`	Converted to the ~ character.
`~a`	abbreviated weekday name
`~A`	full weekday name
`~b`	abbreviated month name
`~B`	full month name
`~c`	time and date using the time and date representation for the locale (`~X ~x`)
`~d`	day of the month as a decimal number (01-31)
`~H`	hour based on a 24-hour clock as a decimal number (00-23)
`~I`	hour based on a 12-hour clock as a decimal number (01-12)
`~j`	day of the year as a decimal number (001-366)
`~m`	month as a decimal number (01-12)
`~M`	minute as a decimal number (00-59)
`~p`	AM/PM designation associated with a 12-hour clock
`~S`	second as a decimal number (00-61)
`~U`	week number of the year; Sunday is first day of week (00-53)
`~w`	weekday as a decimal number (0-6), where Sunday is 0
`~W`	week number of the year; Monday is first day of week (00-53)
`~x`	date using the date representation for the locale
`~X`	time using the time representation for the locale
`~y`	year without century (00-99)
`~Y`	year with century (*e.g.*1990)
`~Z`	time zone name or abbreviation, or no characters if no time zone is determinable

Figure 3.1: `format-date` conversion specifiers

- **year, month, month-day, year-day ⇒ week-day**
  If either of the above rules is able to determine what day it is, the `week-day` field is then set.
- **tz-secs ⇒ tz-name**
  If `tz-secs` is defined, but `tz-name` is not, it is assigned a time-zone name of the form "`UTC+`*`hh`*`:`*`mm`*`:`*`ss`*"; the trailing `:`*`mm`*`:`*`ss`* portion is deleted if it is zeroes.
- **tz-name, date, summer? ⇒ tz-secs, summer?**
  If the date information is provided up to second resolution, `tz-name` is also provided, and `tz-secs` is not set, then `tz-secs` and `summer?` are set to their correct values. Summer-time ambiguities are resolved using the original value of `summer?`. If the time zone doesn't have a summer time variant, then `summer?` is set to `#f`.
- **local time, date, summer? ⇒ tz-name, tz-secs, summer?**
  If the date information is provided up to second resolution, but no time zone information is provided (both `tz-name` and `tz-secs` aren't set), then we proceed as in the above case, except the system's current time zone is used.

These rules allow one particular ambiguity to escape: if both `tz-name` and `tz-secs` are set, they are not brought into agreement. It isn't clear how to do this, nor is it clear which one should take precedence.

Oops: `fill-in-date!` isn't implemented yet.

3.11 Environment variables

(`setenv` *var val*) ⟶ *undefined* procedure
(`getenv` *var*) ⟶ *string* procedure

These functions get and set the process environment, stored in the external C variable `char **environ`. An environment variable *var* is a string. If an environment variable is set to a string *val*, then the process' global environment structure is altered with an entry of the form "*var*=*val*". If *val* is `#f`, then any entry for *var* is deleted.

(`env->alist`) ⟶ *string→string alist* procedure

The `env->alist` procedure converts the entire environment into an alist, *e.g.*,

```
(("TERM" . "vt100")
 ("SHELL" . "/bin/csh")
 ("EDITOR" . "emacs")
 ...)
```

`(alist->env` *alist*`)` $\longrightarrow$ *undefined* procedure

Alist must be an alist whose keys are all strings, and whose values are all either strings or string lists. String lists are converted to colon lists (see below). The alist is installed as the current Unix environment (*i.e.*, converted to a null-terminated C vector of `"`*var=val*`"` strings which is assigned to the global `char **environ`).

The following three functions help the programmer manipulate alist tables in some generally useful ways. They are all defined using `equal?` for key comparison.

`(alist-delete` *key alist*`)` $\longrightarrow$ *alist* procedure

Delete any entry labelled by value *key*.

`(alist-update` *key val alist*`)` $\longrightarrow$ *alist* procedure

Delete *key* from *alist*, then cons on a `(`*key* `.` *val*`)` entry.

`(alist-compress` *alist*`)` $\longrightarrow$ *alist* procedure

Compresses *alist* by removing shadowed entries. Example:

```
;;; Shadowed (1 . c) entry removed.
(alist-compress '( (1 . a) (2 . b) (1 . c) (3 . d) ))
    ==>  ((1 . a) (2 . b) (3 . d))
```

`(with-env*` *env-alist-delta thunk*`)` $\longrightarrow$ *value(s) of thunk* procedure
`(with-total-env*` *env-alist thunk*`)` $\longrightarrow$ *value(s) of thunk* procedure

These procedures call *thunk* in the context of an altered environment. They return whatever values *thunk* returns. Non-local returns restore the environment to its outer value; throwing back into the thunk by invoking a stored continuation restores the environment back to its inner value.

The *env-alist-delta* argument specifies a *modification* to the current environment—*thunk*'s environment is the original environment overridden with the bindings specified by the alist delta.

The *env-alist* argument specifies a complete environment that is installed for *thunk*.

`(with-env` *env-alist-delta . body*`)` $\longrightarrow$ *value(s) of body* syntax
`(with-total-env` *env-alist . body*`)` $\longrightarrow$ *value(s) of body* syntax

These special forms provide syntactic sugar for `with-env*` and `with-total-env*`. The env alists are not evaluated positions, but are implicitly backquoted. In this way, they tend to resemble binding lists for `let` and `let*` forms.

Example: These four pieces of code all run the mailer with special `$TERM` and `$EDITOR` values.

```
(with-env (("TERM" . "xterm") ("EDITOR" . ,my-editor))
  (run (mail shivers@lcs.mit.edu)))

(with-env* `(("TERM" . "xterm") ("EDITOR" . ,my-editor))
  (λ () (run (mail shivers@csd.hku.hk))))

(run (begin (setenv "TERM" "xterm")      ; Env mutation happens
            (setenv "EDITOR" my-editor) ; in the subshell.
            (exec-epf (mail shivers@research.att.com))))

;; In this example, we compute an alternate environment ENV2
;; as an alist, and install it with an explicit call to the
;; EXEC-PATH/ENV procedure.
(let* ((env (env->alist))           ; Get the current environment,
       (env1 (alist-update env  "TERM" "xterm"))      ; and compute
       (env2 (alist-update env1 "EDITOR" my-editor))) ; the new env.
  (run (begin (exec-path/env "mail" env2 "shivers@cs.cmu.edu"))))
```

3.11.1 Path lists and colon lists

Environment variables such as `$PATH` encode a list of strings by separating the list elements with colon delimiters. Once parsed into actual lists, these ordered lists can be manipulated with the following two functions. To convert between the colon-separated string encoding and the list-of-strings representation, see the `field-reader` and `join-strings` functions in section 8.1.

> *Remark:* An earlier release of scsh provided the `split-colon-list` and `string-list->colon-list` functions. These have been removed from scsh, and are replaced by the more general parsers and unparsers of the field-reader module.

(`add-before` *elt before list*) ⟶ *list* — procedure
(`add-after` *elt after list*) ⟶ *list* — procedure

These functions are for modifying search-path lists, where element order is significant.

`add-before` adds *elt* to the list immediately before the first occurrence of *before* in the list. If *before* is not in the list, *elt* is added to the end of the list.

`add-after` is similar: *elt* is added after the last occurrence of *after*. If *after* is not found, *elt* is added to the beginning of the list.

Neither function destructively alters the original path-list. The result may share structure with the original list. Both functions use `equal?` for comparing elements.

3.11.2 `$USER`, `$HOME`, and `$PATH`

Like sh and unlike csh, scsh has *no* interactive dependencies on environment variables. It does, however, initialise certain internal values at startup time from the initial process environment, in particular `$HOME` and `$PATH`. Scsh never uses `$USER` at all. It computes `(user-login-name)` from the system call `(user-uid)`.

`home-directory` — *string*
`exec-path-list` — *string* list

Scsh accesses `$HOME` at start-up time, and stores the value in the global variable `home-directory`. It uses this value for `~` lookups and for returning to home on `(chdir)`.

Scsh accesses `$PATH` at start-up time, colon-splits the path list, and stores the value in the global variable `exec-path-list`. This list is used for `exec-path` and `exec-path/env` searches.

3.12 Terminal device control

Scsh provides a complete set of routines for manipulating terminal devices—putting them in "raw" mode, changing and querying their special characters, modifying their i/o speeds, and so forth. The scsh interface is designed both for generality and portability across different Unix platforms, so you don't have to rewrite your program each time you move to a new system. We've also made an effort to use reasonable, Scheme-like names for the multitudinous named constants involved, so when you are reading code, you'll have less likelihood of getting lost in a bewildering maze of obfuscatory constants named `ICRNL`, `INPCK`, `IUCLC`, and `ONOCR`.

This section can only lay out the basic functionality of the terminal device interface. For further details, see the termios(3) man page on your system, or consult one of the standard Unix texts.

3.12.1 Portability across OS variants

Terminal-control software is inescapably complex, ugly, and low-level. Unix variants each provide their own way of controlling terminal devices, making it difficult to provide interfaces that are portable across different Unix systems. Scsh's terminal support is based primarily upon the POSIX termios interface. Programs that can be written using only the POSIX interface are likely to be widely portable.

The bulk of the documentation that follows consists of several pages worth of tables defining different named constants that enable and disable different features of the terminal driver. Some of these flags are POSIX; others are taken from the two common branches of Unix development, SVR4 and 4.3+ Berkeley. Scsh guarantees that the non-POSIX constants will be bound identifiers.

- If your OS supports a particular non-POSIX flag, its named constant will be bound to the flag's value.
- If your OS doesn't support the flag, its named constant will be present, but bound to `#f`.

This means that if you want to use SVR4 or Berkeley features in a program, your program can portably test the values of the flags before using them—the flags can reliably be referenced without producing OS-dependent "unbound variable" errors.

Finally, note that although POSIX, SVR4, and Berkeley cover the lion's share of the terminal-driver functionality, each operating system inevitably has non-standard extensions. While a particular scsh implementation may provide these extensions, they are not portable, and so are not documented here.

3.12.2 Miscellaneous procedures

(`tty?` *fd/port*) ⟶ *boolean* procedure

Return true if the argument is a tty.

(`tty-file-name` *fd/port*) ⟶ *string* procedure

The argument *fd/port* must be a file descriptor or port open on a tty. Return the file-name of the tty.

3.12.3 The tty-info record type

The primary data-structure that describes a terminal's mode is a `tty-info` record, defined as follows:

```
(define-record tty-info
  control-chars  ; String: Magic input chars
  input-flags    ; Int: Input processing
  output-flags   ; Int: Output processing
  control-flags  ; Int: Serial-line control
  local-flags    ; Int: Line-editting UI
  input-speed    ; Int: Code for input speed
  output-speed   ; Int: Code for output speed
  min            ; Int: Raw-mode input policy
  time)          ; Int: Raw-mode input policy
```

The control-characters string

The `control-chars` field is a character string; its characters may be indexed by integer values taken from table 3.4.

As discussed above, only the POSIX entries in table 3.4 are guaranteed to be legal, integer indices. A program can reliably test the OS to see if the non-POSIX characters are supported by checking the index constants. If the control-character function is supported by the terminal driver, then the corresponding index will be bound to an integer; if it is not supported, the index will be bound to `#f`.

To disable a given control-character function, set its corresponding entry in the `tty-info:control-chars` string to the special character `disable-tty-char` (and then use the (`set-tty-info` *fd/port info*) procedure to update the terminal's state).

The flag fields

The `tty-info` record's `input-flags`, `output-flags`, `control-flags`, and `local-flags` fields are all bit sets represented as two's-complement integers. Their values are composed by or'ing together values taken from the named constants listed in tables 3.5 through 3.9.

As discussed above, only the POSIX entries listed in these tables are guaranteed to be legal, integer flag values. A program can reliably test the OS to see if the non-POSIX flags are supported by checking the named constants. If the feature is supported by the terminal driver, then the corresponding flag will be bound to an integer; if it is not supported, the flag will be bound to `#f`.

The speed fields

The `input-speed` and `output-speed` fields determine the I/O rate of the terminal's line. The value of these fields is an integer giving the speed in bits-per-second. The following speeds are supported by POSIX:

0	134	600	4800
50	150	1200	9600
75	200	1800	19200
110	300	2400	38400

Your OS may accept others; it may also allow the special symbols `'exta` and `'extb`.

The min and time fields

The integer `min` and `time` fields determine input blocking behaviour during non-canonical (raw) input; otherwise, they are ignored. See the termios(3) man page for further details.

Be warned that POSIX allows the base system call's representation of the `tty-info` record to share storage for the `min` field and the `ttychar/eof` element of the control-characters string, and for the `time` field and the `ttychar/eol` element of the control-characters string. Many implementations in fact do this.

To stay out of trouble, set the `min` and `time` fields only if you are putting the terminal into raw mode; set the eof and eol control-characters only if you are putting the terminal into canonical mode. It's ugly, but it's Unix.

3.12.4 Using tty-info records

(`make-tty-info` *if of cf lf ispeed ospeed min time*) $\longrightarrow$ *tty-info-record* procedure
(`copy-tty-info` *tty-info-record*) $\longrightarrow$ *tty-info-record* procedure

> These procedures make it possible to create new `tty-info` records. The typical method for creating a new record is to copy one retrieved by a call to the `tty-info` procedure, then modify the copy as desired. Note that the `make-tty-info` procedure does not take a parameter to define the new record's control characters.[4] Instead, it simply returns a `tty-info` record whose control-character string has all elements initialised to ASCII nul. You may then install the special characters by assigning to the string. Similarly, the control-character string in the record produced by `copy-tty-info` does not share structure with the string in the record being copied, so you may mutate it freely.

[4] Why? Because the length of the string varies from Unix to Unix. For example, the word-erase control character (typically control-w) is provided by most Unixes, but not part of the POSIX spec.

(`tty-info` *[fd/port/fname]*) ⟶ *tty-info-record* procedure

The *fd/port/fname* parameter is an integer file descriptor or Scheme I/O port opened on a terminal device, or a file-name for a terminal device; it defaults to the current input port. This procedure returns a `tty-info` record describing the terminal's current mode.

(`set-tty-info/now` *fd/port/fname info*) ⟶ *no-value* procedure
(`set-tty-info/drain` *fd/port/fname info*) ⟶ *no-value* procedure
(`set-tty-info/flush` *fd/port/fname info*) ⟶ *no-value* procedure

The *fd/port/fname* parameter is an integer file descriptor or Scheme I/O port opened on a terminal device, or a file-name for a terminal device. The procedure chosen determines when and how the terminal's mode is altered:

Procedure	Meaning
`set-tty-info/now`	Make change immediately.
`set-tty-info/drain`	Drain output, then change.
`set-tty-info/flush`	Drain output, flush input, then change.

Oops: If I had defined these with the parameters in the reverse order, I could have made *fd/port/fname* optional. Too late now.

3.12.5 Other terminal-device procedures

(`send-tty-break` *[fd/port/fname duration]*) ⟶ *no-value* procedure

The *fd/port/fname* parameter is an integer file descriptor or Scheme I/O port opened on a terminal device, or a file-name for a terminal device; it defaults to the current output port. Send a break signal to the designated terminal. A break signal is a sequence of continuous zeros on the terminal's transmission line.

The *duration* argument determines the length of the break signal. A zero value (the default) causes a break of between 0.25 and 0.5 seconds to be sent; other values determine a period in a manner that will depend upon local community standards.

(`drain-tty` *[fd/port/fname]*) ⟶ *no-value* procedure

The *fd/port/fname* parameter is an integer file descriptor or Scheme I/O port opened on a terminal device, or a file-name for a terminal device; it defaults to the current output port.

This procedure waits until all the output written to the terminal device has been transmitted to the device. If *fd/port/fname* is an output port with buffered I/O enabled, then the port's buffered characters are flushed before waiting for the device to drain.

(`flush-tty/input` *[fd/port/fname]*) ⟶ *no-value* procedure
(`flush-tty/output` *[fd/port/fname]*) ⟶ *no-value* procedure
(`flush-tty/both` *[fd/port/fname]*) ⟶ *no-value* procedure

The *fd/port/fname* parameter is an integer file descriptor or Scheme I/O port opened on a terminal device, or a file-name for a terminal device; it defaults to the current input port (`flush-tty/input` and `flush-tty/both`), or output port (`flush-tty/output`).

These procedures discard the unread input chars or unwritten output chars in the tty's kernel buffers.

(`start-tty-output` *[fd/port/fname]*) ⟶ *no-value* procedure
(`stop-tty-output` *[fd/port/fname]*) ⟶ *no-value* procedure
(`start-tty-input` *[fd/port/fname]*) ⟶ *no-value* procedure
(`stop-tty-input` *[fd/port/fname]*) ⟶ *no-value* procedure

These procedures can be used to control a terminal's input and output flow. The *fd/port/fname* parameter is an integer file descriptor or Scheme I/O port opened on a terminal device, or a file-name for a terminal device; it defaults to the current input or output port.

The `stop-tty-output` and `start-tty-output` procedures suspend and resume output from a terminal device. The `stop-tty-input` and `start-tty-input` procedures transmit the special STOP and START characters to the terminal with the intention of stopping and starting terminal input flow.

3.12.6 Control terminals, sessions, and terminal process groups

(`open-control-tty` *tty-name [flags]*) ⟶ *port* procedure

This procedure opens terminal device *tty-name* as the process' control terminal (see the `termios` man page for more information on control terminals). The *tty-name* argument is a file-name such as `/dev/ttya`. The *flags* argument is a value suitable as the second argument to the `open-file` call; it defaults to `open/read+write`, causing the terminal to be opened for both input and output.

The port returned is an input port if the *flags* permit it, otherwise an output port. R4RS/Scheme 48/scsh do not have input/output ports, so it's one or the other. However, you can get both read and write ports open on a terminal by opening it read/write, taking the result input port, and duping it to an output port with `dup->outport`.

This procedure guarantees to make the opened terminal the process' control terminal only if the process does not have an assigned control terminal at the time of the call. If the scsh process already has a control terminal, the results are undefined.

To arrange for the process to have no control terminal prior to calling this procedure, use the `become-session-leader` procedure.

`(become-session-leader)` ⟶ *integer* procedure

This is the C `setsid()` call. POSIX job-control has a three-level hierarchy: session/process-group/process. Every session has an associated control terminal. This procedure places the current process into a brand new session, and disassociates the process from any previous control terminal. You may subsequently use `open-control-tty` to open a new control terminal.

It is an error to call this procedure if the current process is already a process-group leader. One way to guarantee this is not the case is only to call this procedure after forking.

`(tty-process-group` *fd/port/fname*`)` ⟶ *integer* procedure
`(set-tty-process-group` *fd/port/fname pgrp*`)` ⟶ *undefined* procedure

This pair of procedures gets and sets the process group of a given terminal.

`(control-tty-file-name)` ⟶ *string* procedure

Return the file-name of the process' control tty. On every version of Unix of which we are aware, this is just the string `"/dev/tty"`. However, this procedure uses the official Posix interface, so it is more portable than simply using a constant string.

3.12.7 Pseudo-terminals

Scsh implements an interface to Berkeley-style pseudo-terminals.

`(fork-pty-session` *thunk*`)` ⟶ *[process pty-in pty-out tty-name]* procedure

This procedure gives a convenient high-level interface to pseudo-terminals. It first allocates a pty/tty pair of devices, and then forks a child to execute procedure *thunk*. In the child process

- Stdio and the current I/O ports are bound to the terminal device.
- The child is placed in its own, new session (see `become-session-leader`).
- The terminal device becomes the new session's controlling terminal (see `open-control-tty`).
- The `(error-output-port)` is unbuffered.

The `fork-pty-session` procedure returns four values: the child's process object, two ports open on the controlling pty device, and the name of the child's corresponding terminal device.

`(open-pty)` $\longrightarrow$ *pty-inport tty-name* procedure

This procedure finds a free pty/tty pair, and opens the pty device with read/write access. It returns a port on the pty, and the name of the corresponding terminal device.

The port returned is an input port—Scheme doesn't allow input/output ports. However, you can easily use `(dup->outport` *pty-inport*`)` to produce a matching output port. You may wish to turn off I/O buffering for this output port.

`(pty-name->tty-name` *pty-name*`)` $\longrightarrow$ *tty-name* procedure
`(tty-name->pty-name` *tty-name*`)` $\longrightarrow$ *pty-name* procedure

These two procedures map between corresponding terminal and pty controller names. For example,

```
(pty-name->tty-name "/dev/ptyq3") ==> "/dev/ttyq3"
(tty-name->pty-name "/dev/ttyrc") ==> "/dev/ptyrc"
```

Remark: This is rather Berkeley-specific. SVR4 ptys are rare enough that I've no real idea if it generalises across the Unix gap. Experts are invited to advise. Users feel free to not worry—the predominance of current popular Unix systems use Berkeley ptys.

`(make-pty-generator)` $\longrightarrow$ *procedure* procedure

This procedure returns a generator of candidate pty names. Each time the returned procedure is called, it produces a new candidate. Software that wishes to search through the set of available ptys can use a pty generator to iterate over them. After producing all the possible ptys, a generator returns `#f` every time it is called. Example:

```
(define pg (make-pty-generator))
(pg) ==> "/dev/ptyp0"
(pg) ==> "/dev/ptyp1"
         ⋮
(pg) ==> "/dev/ptyqe"
(pg) ==> "/dev/ptyqf"   (Last one)
(pg) ==> #f
(pg) ==> #f
         ⋮
```

Scsh	C	Typical char
POSIX		
`ttychar/delete-char`	`ERASE`	del
`ttychar/delete-line`	`KILL`	`^U`
`ttychar/eof`	`EOF`	`^D`
`ttychar/eol`	`EOL`	
`ttychar/interrupt`	`INTR`	`^C`
`ttychar/quit`	`QUIT`	`^\`
`ttychar/suspend`	`SUSP`	`^Z`
`ttychar/start`	`START`	`^Q`
`ttychar/stop`	`STOP`	`^S`
SVR4 and 4.3+BSD		
`ttychar/delayed-suspend`	`DSUSP`	`^Y`
`ttychar/delete-word`	`WERASE`	`^W`
`ttychar/discard`	`DISCARD`	`^O`
`ttychar/eol2`	`EOL2`	
`ttychar/literal-next`	`LNEXT`	`^V`
`ttychar/reprint`	`REPRINT`	`^R`
4.3+BSD		
`ttychar/status`	`STATUS`	`^T`

Table 3.4: Indices into the `tty-info` record's *control-chars* string, and the character traditionally found at each index. Only the indices for the POSIX entries are guaranteed to be non-`#f`.

Scsh	C	Meaning
POSIX		
`ttyin/check-parity`	`INPCK`	Check parity.
`ttyin/ignore-bad-parity-chars`	`IGNPAR`	Ignore chars with parity errors.
`ttyin/mark-parity-errors`	`PARMRK`	Insert chars to mark parity errors.
`ttyin/ignore-break`	`IGNBRK`	Ignore breaks.
`ttyin/interrupt-on-break`	`BRKINT`	Signal on breaks.
`ttyin/7bits`	`ISTRIP`	Strip char to seven bits.
`ttyin/cr->nl`	`ICRNL`	Map carriage-return to newline.
`ttyin/ignore-cr`	`IGNCR`	Ignore carriage-returns.
`ttyin/nl->cr`	`INLCR`	Map newline to carriage-return.
`ttyin/input-flow-ctl`	`IXOFF`	Enable input flow control.
`ttyin/output-flow-ctl`	`IXON`	Enable output flow control.
SVR4 and 4.3+BSD		
`ttyin/xon-any`	`IXANY`	Any char restarts after stop.
`ttyin/beep-on-overflow`	`IMAXBEL`	Ring bell when queue full.
SVR4		
`ttyin/lowercase`	`IUCLC`	Map upper case to lower case.

Table 3.5: Input-flags. These are the named flags for the `tty-info` record's *input-flags* field. These flags generally control the processing of input chars. Only the POSIX entries are guaranteed to be non-`#f`.

Scsh	C	Meaning
POSIX		
`ttyout/enable`	`OPOST`	Enable output processing.
SVR4 and 4.3+BSD		
`ttyout/nl->crnl`	`ONLCR`	Map nl to cr-nl.
4.3+BSD		
`ttyout/discard-eot`	`ONOEOT`	Discard EOT chars.
`ttyout/expand-tabs`	`OXTABS`[5]	Expand tabs.
SVR4		
`ttyout/cr->nl`	`OCRNL`	Map cr to nl.
`ttyout/nl-does-cr`	`ONLRET`	Nl performs cr as well.
`ttyout/no-col0-cr`	`ONOCR`	No cr output in column 0.
`ttyout/delay-w/fill-char`	`OFILL`	Send fill char to delay.
`ttyout/fill-w/del`	`OFDEL`	Fill char is ASCII DEL.
`ttyout/uppercase`	`OLCUC`	Map lower to upper case.

Table 3.6: Output-flags. These are the named flags for the `tty-info` record's *output-flags* field. These flags generally control the processing of output chars. Only the POSIX entries are guaranteed to be non-`#f`.

	Value	Comment
Backspace delay	`ttyout/bs-delay`	Bit-field mask
	`ttyout/bs-delay0`	
	`ttyout/bs-delay1`	
Carriage-return delay	`ttyout/cr-delay`	Bit-field mask
	`ttyout/cr-delay0`	
	`ttyout/cr-delay1`	
	`ttyout/cr-delay2`	
	`ttyout/cr-delay3`	
Form-feed delay	`ttyout/ff-delay`	Bit-field mask
	`ttyout/ff-delay0`	
	`ttyout/ff-delay1`	
Horizontal-tab delay	`ttyout/tab-delay`	Bit-field mask
	`ttyout/tab-delay0`	
	`ttyout/tab-delay1`	
	`ttyout/tab-delay2`	
	`ttyout/tab-delayx`	Expand tabs
Newline delay	`ttyout/nl-delay`	Bit-field mask
	`ttyout/nl-delay0`	
	`ttyout/nl-delay1`	
Vertical tab delay	`ttyout/vtab-delay`	Bit-field mask
	`ttyout/vtab-delay0`	
	`ttyout/vtab-delay1`	
All	`ttyout/all-delay`	Total bit-field mask

Table 3.7: Delay constants. These are the named flags for the `tty-info` record's *output-flags* field. These flags control the output delays associated with printing special characters. They are non-POSIX, and have non-`#f` values only on SVR4 systems.

Scsh	C	Meaning
POSIX		
`ttyc/char-size`	`CSIZE`	Character size mask
`ttyc/char-size5`	`CS5`	5 bits
`ttyc/char-size6`	`CS6`	6 bits
`ttyc/char-size7`	`CS7`	7 bits
`ttyc/char-size8`	`CS8`	8 bits
`ttyc/enable-parity`	`PARENB`	Generate and detect parity.
`ttyc/odd-parity`	`PARODD`	Odd parity.
`ttyc/enable-read`	`CREAD`	Enable reception of chars.
`ttyc/hup-on-close`	`HUPCL`	Hang up on last close.
`ttyc/no-modem-sync`	`LOCAL`	Ignore modem lines.
`ttyc/2-stop-bits`	`CSTOPB`	Send two stop bits.
4.3+BSD		
`ttyc/ignore-flags`	`CIGNORE`	Ignore control flags.
`ttyc/CTS-output-flow-ctl`	`CCTS_OFLOW`	CTS flow control of output
`ttyc/RTS-input-flow-ctl`	`CRTS_IFLOW`	RTS flow control of input
`ttyc/carrier-flow-ctl`	`MDMBUF`	

Table 3.8: Control-flags. These are the named flags for the `tty-info` record's *control-flags* field. These flags generally control the details of the terminal's serial line. Only the POSIX entries are guaranteed to be non-`#f`.

Scsh	C	Meaning
POSIX		
`ttyl/canonical`	`ICANON`	Canonical input processing.
`ttyl/echo`	`ECHO`	Enable echoing.
`ttyl/echo-delete-line`	`ECHOK`	Echo newline after line kill.
`ttyl/echo-nl`	`ECHONL`	Echo newline even if echo is off.
`ttyl/visual-delete`	`ECHOE`	Visually erase chars.
`ttyl/enable-signals`	`ISIG`	Enable ^C, ^Z signalling.
`ttyl/extended`	`IEXTEN`	Enable extensions.
`ttyl/no-flush-on-interrupt`	`NOFLSH`	Don't flush after interrupt.
`ttyl/ttou-signal`	`ITOSTOP`	`SIGTTOU` on background output.
SVR4 and 4.3+BSD		
`ttyl/echo-ctl`	`ECHOCTL`	Echo control chars as "`^X`".
`ttyl/flush-output`	`FLUSHO`	Output is being flushed.
`ttyl/hardcopy-delete`	`ECHOPRT`	Visual erase for hardcopy.
`ttyl/reprint-unread-chars`	`PENDIN`	Retype pending input.
`ttyl/visual-delete-line`	`ECHOKE`	Visually erase a line-kill.
4.3+BSD		
`ttyl/alt-delete-word`	`ALTWERASE`	Alternate word erase algorithm
`ttyl/no-kernel-status`	`NOKERNINFO`	No kernel status on `^T`.
SVR4		
`ttyl/case-map`	`XCASE`	Canonical case presentation

Table 3.9: Local-flags. These are the named flags for the `tty-info` record's *local-flags* field. These flags generally control the details of the line-editting user interface. Only the POSIX entries are guaranteed to be non-`#f`.

Chapter 4

Networking

The Scheme Shell provides a BSD-style sockets interface. There is not an official standard for a network interface for scsh to adopt (this is the subject of the forthcoming Posix.8 standard). However, Berkeley sockets are a *de facto* standard, being found on most Unix workstations and PC operating systems.

It is fairly straightforward to add higher-level network protocols such as smtp, telnet, or http on top of the the basic socket-level support scsh provides. The Scheme Underground has also released a network library with many of these protocols as a companion to the current release of scsh. See this code for examples showing the use of the sockets interface.

4.1 High-level interface

For convenience, and to avoid some of the messy details of the socket interface, we provide a high level socket interface. These routines attempt to make it easy to write simple clients and servers without having to think of many of the details of initiating socket connections. We welcome suggested improvements to this interface, including better names, which right now are solely descriptions of the procedure's action. This might be fine for people who already understand sockets, but does not help the new networking programmer.

(`socket-connect` *protocol-family socket-type . args*) $\longrightarrow$ *socket* procedure

> `socket-connect` is intended for creating client applications. *protocol-family* is specified as either the `protocol-family/internet` or `protocol-family/unix`. *socket-type* is specified as either `socket-type/stream` or `socket-type/datagram`. See `socket` for a more complete description of these terms.

The variable *args* list is meant to specify protocol family specific information. For Internet sockets, this consists of two arguments: a host name and a port number. For Unix sockets, this consists of a pathname.

`socket-connect` returns a `socket` which can be used for input and output from a remote server. See `socket` for a description of the *socket record*.

(`bind-listen-accept-loop` *protocol-family proc arg*) $\longrightarrow$ *does-not-return* procedure

`bind-listen-accept-loop` is intended for creating server applications. *protocol-family* is specified as either the `protocol-family/internet` or `protocol-family/unix`. *proc* is a procedure of two arguments: a `socket` and a socket-address. *arg* specifies a port number for Internet sockets or a pathname for Unix sockets. See `socket` for a more complete description of these terms.

proc is called with a socket and a socket address each time there is a connection from a client application. The socket allows communications with the client. The socket address specifies the address of the remote client.

This procedure does not return, but loops indefinitely accepting connections from client programs.

4.2 Sockets

(`create-socket` *protocol-family type [protocol]*) $\longrightarrow$ *socket* procedure
(`create-socket-pair` *type*) $\longrightarrow$ *[socket$_1$ socket$_2$]* procedure
(`close-socket` *socket*) $\longrightarrow$ *undefined* procedure

A socket is one end of a network connection. Three specific properties of sockets are specified at creation time: the protocol-family, type, and protocol.

The *protocol-family* specifies the protocol family to be used with the socket. This also determines the address family of socket addresses, which are described in more detail below. Scsh currently supports the Unix internal protocols and the Internet protocols using the following constants:

```
protocol-family/unspecified
protocol-family/unix
protocol-family/internet
```

The *type* specifies the style of communication. Examples that your operating system probably provides are stream and datagram sockets. Others maybe available depending on your system. Typical values are:

```
socket-type/stream
socket-type/datagram
socket-type/raw
```

The *protocol* specifies a particular protocol to use within a protocol family and type. Usually only one choice exists, but it's probably safest to set this explicitly. See the protocol database routines for information on looking up protocol constants.

New sockets are typically created with `create-socket`. However, `create-socket-pair` can also be used to create a pair of connected sockets in the `protocol-family/unix` protocol-family. The value of a returned socket is a *socket record*, defined to have the following structure:

```
(define-record socket
  family                                  ; protocol family
  inport                                  ; input-port
  outport)                                ; output-port
```

The `family` specifies the protocol family of the socket. The `inport` and `outport` fields are ports that can be used for input and output, respectively. For a stream socket, they are only usable after a connection has been established via `connect-socket` or `accept-connection`. For a datagram socket, *outport* can be immediately using `send-message`, and *inport* can be used after `bind` has created a local address.

`close-socket` provides a convenient way to close a socket's port. It is preferred to explicitly closing the inport and outport because using `close` on sockets is not currently portable across operating systems.

4.3 Socket addresses

The format of a socket-address depends on the address family of the socket. Address-family-specific routines are provided to convert protocol-specific addresses to socket addresses. The value returned by these routines is a *socket-address record*, defined to have the following visible structure:

```
(define-record socket-address
  family)                                 ; address family
```

The `family` is one of the following constants:

```
address-family/unspecified
address-family/unix
address-family/internet
```

(unix-address->socket-address *pathname*) ⟶ *socket-address* procedure

unix-address->socket-address returns a *socket-address* based on the string *pathname*. There is a system dependent limit on the length of *pathname*.

(internet-address->socket-address *host-address service-port*) ⟶ *socket-address* procedure

internet-address->socket-address returns a *socket-address* based on an integer *host-address* and an integer *service-port*. Besides being a 32-bit host address, an Internet host address can also be one of the following constants:

```
internet-address/any
internet-address/loopback
internet-address/broadcast
```

The use of internet-address/any is described below in bind-socket. internet-address/loopback is an address that always specifies the local machine. internet-address/broadcast is used for network broadcast communications.

For information on obtaining a host's address, see the host-info function.

(socket-address->unix-address *socket-address*) ⟶ *pathname* procedure
(socket-address->internet-address *socket-address*) ⟶ *[host-address service-port]* procedure

The routines socket-address->internet-address and socket-address->unix-address return the address-family-specific addresses. Be aware that most implementations don't correctly return anything more than an empty string for addresses in the Unix address-family.

4.4 Socket primitives

The procedures in this section are presented in the order in which a typical program will use them. Consult a text on network systems programming for more information on sockets.[1] The last two tutorials are freely available as part

[1] Some recommended ones are:

- "Unix Network Programming" by W. Richard Stevens
- "An Introductory 4.3BSD Interprocess Communication Tutorial." (reprinted in UNIX Programmer's Supplementary Documents Volume 1, PS1:7)
- "An Advanced 4.3BSD Interprocess Communication Tutorial." (reprinted in UNIX Programmer's Supplementary Documents Volume 1, PS1:8)

of BSD. In the absence of these, your Unix manual pages for socket might be a good starting point for information.

(`connect-socket` *socket socket-address*) ⟶ *undefined* procedure

> `connect-socket` sets up a connection from a *socket* to a remote *socket-address*. A connection has different meanings depending on the socket type. A stream socket must be connected before use. A datagram socket can be connected multiple times, but need not be connected at all if the remote address is specified with each `send-message`, described below. Also, datagram sockets may be disassociated from a remote address by connecting to a null remote address.

(`bind-socket` *socket socket-address*) ⟶ *undefined* procedure

> `bind-socket` assigns a certain local *socket-address* to a *socket*. Binding a socket reserves the local address. To receive connections after binding the socket, use `listen-socket` for stream sockets and `receive-message` for datagram sockets.
>
> Binding an Internet socket with a host address of `internet-address/any` indicates that the caller does not care to specify from which local network interface connections are received. Binding an Internet socket with a service port number of zero indicates that the caller has no preference as to the port number assigned.
>
> Binding a socket in the Unix address family creates a socket special file in the file system that must be deleted before the address can be reused. See `delete-file`.

(`listen-socket` *socket backlog*) ⟶ *undefined* procedure

> `listen-socket` allows a stream *socket* to start receiving connections, allowing a queue of up to *backlog* connection requests. Queued connections may be accepted by `accept-connection`.

(`accept-connection` *socket*) ⟶ *[new-socket socket-address]* procedure

> `accept-connection` receives a connection on a *socket*, returning a new socket that can be used for this connection and the remote socket address associated with the connection.

(`socket-local-address` *socket*) ⟶ *socket-address* procedure
(`socket-remote-address` *socket*) ⟶ *socket-address* procedure

> Sockets can be associated with a local address or a remote address or both. `socket-local-address` returns the local *socket-address* record associated with *socket*. `socket-remote-address` returns the remote *socket-address* record associated with *socket*.

(`shutdown-socket` *socket how-to*) ⟶ *undefined* procedure

`shutdown-socket` shuts down part of a full-duplex socket. The method of shutting done is specified by the *how-to* argument, one of:

```
shutdown/receives
shutdown/sends
shutdown/sends+receives
```

4.5 Performing input and output on sockets

(`receive-message` *socket length [flags]*) ⟶ *[string-or-*`#f`socket-address] procedure
(`receive-message!` *socket string [start] [end] [flags]*) ⟶ *[count-or-*`#f`socket-address] procedure
(`receive-message/partial` *socket length [flags]*) ⟶ *[string-or-*`#f`socket-address] procedure
(`receive-message!/partial` *socket string [start] [end] [flags]*) ⟶ *[count-or-*`#f`socket-address] procedure

(`send-message` *socket string [start] [end] [flags] [socket-address]*) ⟶ *undefined* procedure
(`send-message/partial` *socket string [start] [end] [flags] [socket-address]*) ⟶ *count* procedure

For most uses, standard input and output routines such as `read-string` and `write-string` should suffice. However, in some cases an extended interface is required. The `receive-message` and `send-message` calls parallel the `read-string` and `write-string` calls with a similar naming scheme.

One additional feature of these routines is that `receive-message` returns the remote *socket-address* and *send-message* takes an optional remote `socket-address`. This allows a program to know the source of input from a datagram socket and to use a datagram socket for output without first connecting it.

All of these procedures take an optional *flags* field. This argument is an integer bit-mask, composed by or'ing together the following constants:

```
message/out-of-band
message/peek
message/dont-route
```

See `read-string` and `write-string` for a more detailed description of the arguments and return values.

4.6 Socket options

(`socket-option` *socket level option*) ⟶ *value* procedure
(`set-socket-option` *socket level option value*) ⟶ *undefined* procedure

`socket-option` and `set-socket-option` allow the inspection and modification, respectively, of several options available on sockets. The *level* argument specifies what protocol level is to be examined or affected. A level of `level/socket` specifies the highest possible level that is available on all socket types. A specific protocol number can also be used as provided by `protocol-info`, described below.

There are several different classes of socket options. The first class consists of boolean options which can be either true or false. Examples of this option type are:

```
socket/debug
socket/accept-connect
socket/reuse-address
socket/keep-alive
socket/dont-route
socket/broadcast
socket/use-loop-back
socket/oob-inline
socket/use-privileged
socket/cant-signal
tcp/no-delay
```

Value options are another category of socket options. Options of this type are an integer value. Examples of this option type are:

```
socket/send-buffer
socket/receive-buffer
socket/send-low-water
socket/receive-low-water
socket/error
socket/type
ip/time-to-live
tcp/max-segment
```

A third option type specifies how long for data to linger after a socket has been closed. There is only one option of this type: `socket/linger`. It is set with either `#f`to disable it or an integer number of seconds to linger and returns a value of the same type upon inspection.

The fourth and final option type of this time is a timeout option. There are two examples of this option type: `socket/send-timeout` and `socket/receive-timeout`. These are set with a real number of microseconds resolution and returns a value of the same type upon inspection.

4.7 Database-information entries

(`host-info` *name-or-socket-address*) $\longrightarrow$ *host-info* procedure
(`network-info` *name-or-socket-address*) $\longrightarrow$ *network-info* procedure
(`service-info` *name-or-number [protocol-name]*) $\longrightarrow$ *service-info* procedure
(`protocol-info` *name-or-number*) $\longrightarrow$ *protocol-info* procedure

`host-info` allows a program to look up a host entry based on either its string *name* or *socket-address*. The value returned by this routine is a *host-info record,* defined to have the following structure:

```
(define-record host-info
  name                                      ; Host name
  aliases                                   ; Alternative names
  addresses)                                ; Host addresses
```

`host-info` could fail and raise an error for one of the following reasons:

```
herror/host-not-found
herror/try-again
herror/no-recovery
herror/no-data
herror/no-address
```

`network-info` allows a program to look up a network entry based on either its string *name* or *socket-address*. The value returned by this routine is a *network-info record,* defined to have the following structure:

```
(define-record network-info
  name                                      ; Network name
  aliases                                   ; Alternative names
  net)                                      ; Network number
```

`service-info` allows a program to look up a service entry based on either its string *name* or integer *port*. The value returned by this routine is a *service-info record,* defined to have the following structure:

```
(define-record service-info
  name                                      ; Service name
  aliases                                   ; Alternative names
  port                                      ; Port number
  protocol)                                 ; Protocol name
```

`protocol-info` allows a program to look up a protocol entry based on either its string *name* or integer *number*. The value returned by this routine is a *protocol-info record,* defined to have the following structure:

```
(define-record protocol-info
  name                                    ; Protocol name
  aliases                                 ; Alternative names
  number)                                 ; Protocol number)
```

Chapter 5

Strings and characters

Scsh provides a set of procedures for processing strings and characters. The procedures provided match regular expressions, search strings, parse file-names, and manipulate sets of characters.

Also see chapters 6, 7 and 8 on regular-expressions, record I/O, field parsing, and the awk loop. The procedures documented there allow you to search and pattern-match strings, read character-delimited records from ports, use regular expressions to split the records into fields (for example, splitting a string at every occurrence of colon or white-space), and loop over streams of these records in a convenient way.

5.1 String manipulation

Strings are the basic communication medium for Unix processes, so a shell language must have reasonable facilities for manipulating them.

5.1.1 Manipulating file-names

These procedures do not access the file-system at all; they merely operate on file-name strings. Much of this structure is patterned after the gnu emacs design. Perhaps a more sophisticated system would be better, something like the pathname abstractions of COMMON LISP or MIT Scheme. However, being Unix-specific, we can be a little less general.

Terminology

These procedures carefully adhere to the POSIX standard for file-name resolution, which occasionally entails some slightly odd things. This section will describe these rules, and give some basic terminology.

A *file-name* is either the file-system root ("/"), or a series of slash-terminated directory components, followed by a a file component. Root is the only file-name that may end in slash. Some examples:

File name	Dir components	File component
`src/des/main.c`	`("src" "des")`	`"main.c"`
`/src/des/main.c`	`("" "src" "des")`	`"main.c"`
`main.c`	`()`	`"main.c"`

Note that the relative filename `src/des/main.c` and the absolute filename `/src/des/main.c` are distinguished by the presence of the root component `""` in the absolute path.

Multiple embedded slashes within a path have the same meaning as a single slash. More than two leading slashes at the beginning of a path have the same meaning as a single leading slash—they indicate that the file-name is an absolute one, with the path leading from root. However, POSIX permits the OS to give special meaning to *two* leading slashes. For this reason, the routines in this section do not simplify two leading slashes to a single slash.

A file-name in *directory form* is either a file-name terminated by a slash, *e.g.*, "`/src/des/`", or the empty string, "". The empty string corresponds to the current working directory, whose file-name is dot ("`.`"). Working backwards from the append-a-slash rule, we extend the syntax of POSIX file-names to define the empty string to be a file-name form of the root directory "`/`". (However, "`/`" is also acceptable as a file-name form for root.) So the empty string has two interpretations: as a file-name form, it is the file-system root; as a directory form, it is the current working directory. Slash is also an ambiguous form: `/` is both a directory-form and a file-name form.

The directory form of a file-name is very rarely used. Almost all of the procedures in scsh name directories by giving their file-name form (without the trailing slash), not their directory form. So, you say "`/usr/include`", and "`.`", not "`/usr/include/`" and "". The sole exceptions are `file-name-as-directory` and `directory-as-file-name`, whose jobs are to convert back-and-forth between these forms, and `file-name-directory`, whose job it is to split out the directory portion of a file-name. However, most procedures that expect a directory argument will coerce a file-name in directory form to file-name form if it does not have a trailing slash. Bear in mind that the ambiguous case, empty string, will be interpreted in file-name form, *i.e.*, as root.

Procedures

(`file-name-directory?` *fname*) ⟶ *boolean* — procedure
(`file-name-non-directory?` *fname*) ⟶ *boolean* — procedure

These predicates return true if the string is in directory form, or file-name form (see the above discussion of these two forms). Note that they both return true on the ambiguous case of empty string, which is both a directory (current working directory), and a file name (the file-system root).

File name	`...-directory?`	`...-non-directory?`
`"src/des"`	`#f`	`#t`
`"src/des/"`	`#t`	`#f`
`"/"`	`#t`	`#f`
`"."`	`#f`	`#t`
`""`	`#t`	`#t`

(`file-name-as-directory` *fname*) ⟶ *string* — procedure

Convert a file-name to directory form. Basically, add a trailing slash if needed:

```
(file-name-as-directory "src/des")  ⟹  "src/des/"
(file-name-as-directory "src/des/") ⟹  "src/des/"
```

`.`, `/`, and `""` are special:

```
(file-name-as-directory ".")        ⟹  ""
(file-name-as-directory "/")        ⟹  "/"
(file-name-as-directory "")         ⟹  "/"
```

(`directory-as-file-name` *fname*) ⟶ *string* — procedure

Convert a directory to a simple file-name. Basically, kill a trailing slash if one is present:

```
(directory-as-file-name "foo/bar/") ⟹  "foo/bar"
```

`/` and `""` are special:

```
(directory-as-file-name "/")        ⟹  "/"
(directory-as-file-name "")         ⟹  "." (i.e., the cwd)
```

(`file-name-absolute?` *fname*) ⟶ *boolean* — procedure

Does *fname* begin with a root or `~` component? (Recognising `~` as a home-directory specification is an extension of POSIX rules.)

```
(file-name-absolute? "/usr/shivers")  ⟹  #t
(file-name-absolute? "src/des")       ⟹  #f
(file-name-absolute? "~/src/des")     ⟹  #t
```

Non-obvious case:

```
(file-name-absolute? "")              ⟹  #t (i.e., root)
```

(file-name-directory *fname*) ⟶ *string or false* procedure

Return the directory component of *fname* in directory form. If the file-name is already in directory form, return it as-is.

```
(file-name-directory "/usr/bdc")      ⟹  "/usr/"
(file-name-directory "/usr/bdc/")     ⟹  "/usr/bdc/"
(file-name-directory "bdc/.login")    ⟹  "bdc/"
(file-name-directory "main.c")        ⟹  ""
```

Root has no directory component:

```
(file-name-directory "/")             ⟹  ""
(file-name-directory "")              ⟹  ""
```

(file-name-nondirectory *fname*) ⟶ *string* procedure

Return non-directory component of fname.

```
(file-name-nondirectory "/usr/ian")     ⟹  "ian"
(file-name-nondirectory "/usr/ian/")    ⟹  ""
(file-name-nondirectory "ian/.login")   ⟹  ".login"
(file-name-nondirectory "main.c")       ⟹  "main.c"
(file-name-nondirectory "")             ⟹  ""
(file-name-nondirectory "/")            ⟹  "/"
```

(split-file-name *fname*) ⟶ *string list* procedure

Split a file-name into its components.

```
(split-file-name "src/des/main.c")
   ⟹      ("src" "des" "main.c")

(split-file-name "/src/des/main.c")
   ⟹   ("" "src" "des" "main.c")

(split-file-name "main.c")
   ⟹                  ("main.c")

(split-file-name "/")
   ⟹  ("")
```

`(path-list->file-name` *path-list [dir]*`)` $\longrightarrow$ *string* procedure

Inverse of `split-file-name`.

```
(path-list->file-name '("src" "des" "main.c"))
    ==>  "src/des/main.c"
(path-list->file-name '("" "src" "des" "main.c"))
    ==>  "/src/des/main.c"
```

Optional *dir* arg anchors relative path-lists:

```
(path-list->file-name '("src" "des" "main.c")
                      "/usr/shivers")
    ==>  "/usr/shivers/src/des/main.c"
```

The optional *dir* argument is usefully `(cwd)`.

`(file-name-extension` *fname*`)` $\longrightarrow$ *string* procedure

Return the file-name's extension.

```
(file-name-extension "main.c")        ==>  ".c"
(file-name-extension "main.c.old")    ==>  ".old"
(file-name-extension "/usr/shivers")  ==>  ""
```

Weird cases:

```
(file-name-extension "foo.")   ==>  "."
(file-name-extension "foo..")  ==>  "."
```

Dot files are not extensions:

```
(file-name-extension "/usr/shivers/.login")  ==>  ""
```

`(file-name-sans-extension` *fname*`)` $\longrightarrow$ *string* procedure

Return everything but the extension.

```
(file-name-sans-extension "main.c")      ==>  "main"
(file-name-sans-extension "main.c.old")  ==>  "main.c""
(file-name-sans-extension "/usr/shivers")
    ==>  "/usr/shivers"
```

Weird cases:

```
(file-name-sans-extension "foo.")   ==>  "foo"
(file-name-sans-extension "foo..")  ==>  "foo."
```

Dot files are not extensions:

```
(file-name-sans-extension "/usr/shivers/.login")
    ==>  "/usr/shivers/.login
```

Note that appending the results of `file-name-extension` and `file-name-sans-extension` in all cases produces the original file-name.

`(parse-file-name` *fname*`)` ⟶ *[dir name extension]* procedure

Let *f* be `(file-name-nondirectory` *fname*`)`. This function returns the three values:

- `(file-name-directory` *fname*`)`
- `(file-name-sans-extension` *f*`))`
- `(file-name-extension` *f*`)`

The inverse of `parse-file-name`, in all cases, is `string-append`. The boundary case of `/` was chosen to preserve this inverse.

`(replace-extension` *fname ext*`)` ⟶ *string* procedure

This procedure replaces *fname*'s extension with *ext*. It is exactly equivalent to

`(string-append (file-name-sans-extension` *fname*`)` *ext*`)`

`(simplify-file-name` *fname*`)` ⟶ *string* procedure

Removes leading and internal occurrences of dot. A trailing dot is left alone, as the parent could be a symlink. Removes internal and trailing double-slashes. A leading double-slash is left alone, in accordance with POSIX. However, triple and more leading slashes are reduced to a single slash, in accordance with POSIX. Double-dots (parent directory) are left alone, in case they come after symlinks or appear in a `/../`*machine*`/...` "super-root" form (which POSIX permits).

`(resolve-file-name` *fname [dir]*`)` ⟶ *string* procedure

- Do ˜ expansion.
- If *dir* is given, convert a relative file-name to an absolute file-name, relative to directory *dir*.

`(expand-file-name` *fname [dir]*`)` ⟶ *string* procedure

Resolve and simplify the file-name.

`(absolute-file-name` *fname [dir]*`)` $\longrightarrow$ *string* procedure

Convert file-name *fname* into an absolute file name, relative to directory *dir*, which defaults to the current working directory. The file name is simplified before being returned.

This procedure does not treat a leading tilde character specially.

`(home-dir` *[user]*`)` $\longrightarrow$ *string* procedure

`home-dir` returns *user*'s home directory. *User* defaults to the current user.

```
(home-dir)           ==>  "/user1/lecturer/shivers"
(home-dir "ctkwan")  ==>  "/user0/research/ctkwan"
```

`(home-file` *[user] fname*`)` $\longrightarrow$ *string* procedure

Returns file-name *fname* relative to *user*'s home directory; *user* defaults to the current user.

```
(home-file "man")           ==>  "/usr/shivers/man"
(home-file "fcmlau" "man")  ==>  "/usr/fcmlau/man"
```

The general `substitute-env-vars` string procedure, defined in the previous section, is also frequently useful for expanding file-names.

5.1.2 Other string manipulation facilities

`(index` *string char [start]*`)` $\longrightarrow$ *fixnum or false* procedure
`(rindex` *string char [start]*`)` $\longrightarrow$ *fixnum or false* procedure

These procedures search through *string* looking for an occurrence of character *char*. `index` searches left-to-right; `rindex` searches right-to-left.

`index` returns the smallest index i of *string* greater than or equal to *start* such that *string*$[i] =$ *char*. The default for *start* is zero. If there is no such match, `index` returns false.

`rindex` returns the largest index i of *string* less than *start* such that *string*$[i] =$ *char*. The default for *start* is `(string-length` *string*`)`. If there is no such match, `rindex` returns false.

I should probably snarf all the MIT Scheme string functions, and stick them in a package. Unix programs need to mung character strings a lot.

MIT string match commands:

```
[sub]string-match-forward,backward[-ci]
[sub]string-prefix,suffix[-ci]?
[sub]string-find-next,previous-char[-ci]
[sub]string-find-next,previous-char-in-set
[sub]string-replace[!]
...etc.
```

These are not currently provided.

(`substitute-env-vars` *fname*) $\longrightarrow$ *string* procedure

> Replace occurrences of environment variables with their values. An environment variable is denoted by a dollar sign followed by alphanumeric chars and underscores, or is surrounded by braces.
>
> ```
> (substitute-env-vars "$USER/.login")
>     ⟹  "shivers/.login"
> (substitute-env-vars "${USER}_log")  ⟹  "shivers_log"
> ```

5.2 ASCII encoding

(`char->ascii` *character*) $\longrightarrow$ *integer* procedure
(`ascii->char` *integer*) $\longrightarrow$ *character* procedure

> These are identical to `char->integer` and `integer->char` except that they use the ASCII encoding.

5.3 Character sets

Scsh provides a `char-set` type for expressing sets of characters. These sets are used by some of the delimited-input procedures (section 8.1). Scsh's character set package was adapted and extended from Project Mac's MIT Scheme package. Note that the character type used in the current implementation corresponds to the ASCII character set—but you would be wise not to build this assumption into your code if you can help it.[1]

(`char-set?` *x*) $\longrightarrow$ *boolean* procedure

> Is the object *x* a character set?

[1]Actually, it's slightly uglier than that, albeit somewhat more useful. The current character type corresponds to an eight-bit superset of ASCII. The `ascii->char` and `char->ascii` functions will preserve this eighth bit. However, none of the the high 128 characters appear in any of the standard character sets defined in section 5.3.4, except for `char-set:full`. If someone would email the authors a listing of the full Latin-1 definition, we'll be happy to upgrade these sets' definitions to make them Latin-1 compliant.

(`char-set=` cs_1 cs_2. . .) ⟶ *boolean* procedure

Are the character sets equal?

(`char-set<=` cs_1 cs_2. . .) ⟶ *boolean* procedure

Returns true if every character set cs_i is a subset of character set cs_{i+1}.

(`char-set-fold` *kons knil cs*) ⟶ *object* procedure

This is the fundamental iterator for character sets. Applies the function *kons* across the character set *cs* using initial state value *knil*. That is, if *cs* is the empty set, the procedure returns *knil*. Otherwise, some element *c* of *cs* is chosen; let *cs′* be the remaining, unchosen characters. The procedure returns

(`char-set-fold` *kons* (*kons c knil*) *cs′*)

For example, we could define `char-set-members` (see below) as

```
(lambda (cs) (char-set-fold cons '() cs))
```

Remark: This procedure was formerly named `reduce-char-set` . The old binding is still provided, but is deprecated and will probably vanish in a future release.

(`char-set-for-each` *p cs*) ⟶ *undefined* procedure

Apply procedure *p* to each character in the character set *cs*. Note that the order in which *p* is applied to the characters in the set is not specified, and may even change from application to application.

5.3.1 Creating character sets

(`char-set` $char_1$. . .) ⟶ *char-set* procedure

Return a character set containing the given characters.

(`chars->char-set` *chars*) ⟶ *char-set* procedure

Return a character set containing the characters in the list *chars*.

(`string->char-set` *s*) ⟶ *char-set* procedure

Return a character set containing the characters in the string *s*.

(`predicate->char-set` *pred*) ⟶ *char-set* procedure

Returns a character set containing every character *c* such that (*pred c*) returns true.

(`ascii-range->char-set` *lower upper*) ⟶ *char-set* procedure

Returns a character set containing every character whose ASCII code lies in the half-open range [*lower*, *upper*).

5.3.2 Querying character sets

(`char-set-members` *char-set*) $\longrightarrow$ *character-list* procedure

This procedure returns a list of the members of *char-set*.

(`char-set-contains?` *char-set char*) $\longrightarrow$ *boolean* procedure

This procedure tests *char* for membership in set *char-set*.

> *Remark:* Previous releases of scsh called this procedure `char-set-member?`, reversing the order of the arguments. This made sense, but was unfortunately the reverse order in which the arguments appear in MIT Scheme. A reasonable argument order was not backwards-compatible with MIT Scheme; on the other hand, the MIT Scheme argument order was counter-intuitive and at odds with common mathematical notation and the `member` family of R4RS procedures.
>
> We sought to escape the dilemma by shifting to a new name.

(`char-set-size` *cs*) $\longrightarrow$ *integer* procedure

Returns the number of elements in character set *cs*.

(`char-set-every?` *pred cs*) $\longrightarrow$ *boolean* procedure
(`char-set-any?` *pred cs*) $\longrightarrow$ *object* procedure

The `char-set-every?` procedure returns true if predicate *pred* returns true of every character in the character set *cs*.

Likewise, `char-set-any?` applies *pred* to every character in character set *cs*, and returns the first true value it finds. If no character produces a true value, it returns false.

The order in which these procedures sequence through the elements of *cs* is not specified.

5.3.3 Character-set algebra

(`char-set-invert` *char-set*) $\longrightarrow$ *char-set* procedure
(`char-set-union` $char\text{-}set_1 \ldots$) $\longrightarrow$ *char-set* procedure
(`char-set-intersection` $char\text{-}set_1$ $char\text{-}set_2 \ldots$) $\longrightarrow$ *char-set* procedure
(`char-set-difference` $char\text{-}set_1$ $char\text{-}set_2 \ldots$) $\longrightarrow$ *char-set* procedure

These procedures implement set complement, union, intersection, and difference for character sets. The union, intersection, and difference operations are n-ary, associating to the left; the difference function requires at least one argument, while union and intersection may be applied to zero arguments.

(char-set-adjoin *cs* $char_1 \ldots$) ⟶ *char-set* procedure
(char-set-delete *cs* $char_1 \ldots$) ⟶ *char-set* procedure

Add/delete the $char_i$ characters to/from character set *cs*.

5.3.4 Standard character sets

Several character sets are predefined for convenience:

`char-set:lower-case`	Lower-case alphabetic chars
`char-set:upper-case`	Upper-case alphabetic chars
`char-set:alphabetic`	Alphabetic chars
`char-set:numeric`	Decimal digits: 0–9
`char-set:alphanumeric`	Alphabetic or numeric
`char-set:graphic`	Printing characters except space
`char-set:printing`	Printing characters including space
`char-set:whitespace`	Whitespace characters
`char-set:control`	Control characters
`char-set:punctuation`	Punctuation characters
`char-set:hex-digit`	A hexadecimal digit: 0–9, A–F, a–f
`char-set:blank`	Blank characters
`char-set:ascii`	A character in the ASCII set.
`char-set:empty`	Empty set
`char-set:full`	All characters

The first eleven of these correspond to the character classes defined in Posix. Note that there may be characters in `char-set:alphabetic` that are neither upper or lower case—this might occur in implementations that use a character type richer than ASCII, such as Unicode. A "graphic character" is one that would put ink on your page. While the exact composition of these sets may vary depending upon the character type provided by the Scheme system upon which scsh is running, here are the definitions for some of the sets in an ASCII character set:

char-set:alphabetic	A–Z and a–z
char-set:lower-case	a–z
char-set:upper-case	A–Z
char-set:graphic	Alphanumeric + punctuation
char-set:whitespace	Space, newline, tab, page, vertical tab, carriage return
char-set:blank	Space and tab
char-set:control	ASCII 0–31 and 127
char-set:punctuation	`!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~`

`(char-alphabetic?` *character*`)` ⟶ *boolean* procedure
`(char-lower-case?` *character*`)` ⟶ *boolean* procedure
`(char-upper-case?` *character*`)` ⟶ *boolean* procedure
`(char-numeric?` *character*`)` ⟶ *boolean* procedure
`(char-alphanumeric?` *character*`)` ⟶ *boolean* procedure
`(char-graphic?` *character*`)` ⟶ *boolean* procedure
`(char-printing?` *character*`)` ⟶ *boolean* procedure
`(char-whitespace?` *character*`)` ⟶ *boolean* procedure
`(char-blank?` *character*`)` ⟶ *boolean* procedure
`(char-control?` *character*`)` ⟶ *boolean* procedure
`(char-punctuation?` *character*`)` ⟶ *boolean* procedure
`(char-hex-digit?` *character*`)` ⟶ *boolean* procedure
`(char-ascii?` *character*`)` ⟶ *boolean* procedure

These predicates are defined in terms of the above character sets.

5.3.5 Linear-update character-set operations

These procedures have a hybrid pure-functional/side-effecting semantics: they are allowed, but not required, to side-effect one of their parameters in order to construct their result. An implementation may legally implement these procedures as pure, side-effect-free functions, or it may implement them using side effects, depending upon the details of what is the most efficient or simple to implement in terms of the underlying representation.

What this means is that clients of these procedures *may not* rely upon these procedures working by side effect. For example, this is not guaranteed to work:

```
(let ((cs (char-set #\a #\b #\c)))
  (char-set-adjoin! cs #\d)
  cs) ; Could be either {a,b,c} or {a,b,c,d}.
```

However, this is well-defined:

```
(let ((cs (char-set #\a #\b #\c)))
  (char-set-adjoin! cs #\d)) ; {a,b,c,d}
```

So clients of these procedures write in a functional style, but must additionally be sure that, when the procedure is called, there are no other live pointers to the potentially-modified character set (hence the term "linear update").

There are two benefits to this convention:

- Implementations are free to provide the most efficient possible implementation, either functional or side-effecting.

- Programmers may nonetheless continue to assume that character sets are purely functional data structures: they may be reliably shared without needing to be copied, uniquified, and so forth.

In practice, these procedures are most useful for efficiently constructing character sets in a side-effecting manner, in some limited local context, before passing the character set outside the local construction scope to be used in a functional manner.

Scsh provides no assistance in checking the linearity of the potentially side-effected parameters passed to these functions — there's no linear type checker or run-time mechanism for detecting violations.

`(char-set-copy` *cs*`)` $\longrightarrow$ *char-set* procedure

> Returns a copy of the character set *cs*. "Copy" means that if either the input parameter or the result value of this procedure is passed to one of the linear-update procedures described below, the other character set is guaranteed not to be altered. (A system that provides pure-functional implementations of the rest of the linear-operator suite could implement this procedure as the identity function.)

`(char-set-adjoin!` *cs* $char_1 \ldots$`)` $\longrightarrow$ *char-set* procedure

> Add the $char_i$ characters to character set *cs*, and return the result. This procedure is allowed, but not required, to side-effect *cs*.

`(char-set-delete!` *cs* $char_1 \ldots$`)` $\longrightarrow$ *char-set* procedure

> Remove the $char_i$ characters to character set *cs*, and return the result. This procedure is allowed, but not required, to side-effect *cs*.

`(char-set-invert!` *char-set*`)` $\longrightarrow$ *char-set* procedure
`(char-set-union!` $char\text{-}set_1$ $char\text{-}set_2 \ldots$`)` $\longrightarrow$ *char-set* procedure
`(char-set-intersection!` $char\text{-}set_1$ $char\text{-}set_2 \ldots$`)` $\longrightarrow$ *char-set* procedure
`(char-set-difference!` $char\text{-}set_1$ $char\text{-}set_2 \ldots$`)` $\longrightarrow$ *char-set* procedure

> These procedures implement set complement, union, intersection, and difference for character sets. They are allowed, but not required, to side-effect their first parameter. The union, intersection, and difference operations are n-ary, associating to the left.

Chapter 6

Pattern-matching strings with regular expressions

Scsh provides a rich facility for matching regular-expression patterns in strings. The system is composed of several pieces:

- An s-expression notation for writing down general regular expressions. In most systems, regexp patterns are encoded as string literals, such as `"g(oo|ee)se"`. In scsh, they are written using s-expressions, such as `(: "g" (| "oo" "ee") "se")`, and are called *sre's*. The sre notation has several advantages over the traditional string-based notation. It's more expressive, can be commented, and can be indented to expose the structure of the form.

- An abstract data type (ADT) representation for regexp values. Traditional regular-expression systems compute regular expressions from run-time values using strings. This can be awkward. Scsh, instead, provides a separate data type for regexps, with a set of basic constructor and accessor functions; regular expressions can be dynamically computed and manipulated using these functions.

- Some tools that work on the regexp ADT: case-sensitve to case-insensitive regexp transform, a regexp simplifier, and so forth.

- Parsers and unparsers that can convert between external representations and the regexp ADT. The supported external representations are
  - Posix strings
  - S-expression notation (that is, sre's)

Being able to convert regexps to Posix strings allows implementations to implement regexp matching using standard Posix C-based engines.

- Macro support for the s-expression notation. The `rx` macro provides a new special form that allows you to embed regexps in the s-expression notation within a Scheme program. Evaluating the macro form produces a regexp ADT value which can be used by Scheme pattern-matching procedures and other regexp consumers.

- Pattern-matching and searching procedures. Spencer's Posix regexp engine is linked in to the runtime; the regexp code uses this engine to provide text matching.

The regexp language supported is a complete superset of Posix functionality, providing:

- sequencing and choice (`|`)
- repetition (`*`, `+`, `?`, $\{m\texttt{,}n\}$)
- character classes (*e.g.*, `[aeiou]`) and wildcard (`.`)
- beginning/end of string anchors (`^`, `$`)
- beginning/end of line anchors
- beginning/end of word anchors
- case-sensitivity control
- submatch-marking

6.1 Summary SRE syntax

The following figures give a summary of the SRE syntax; the next section is a friendlier tutorial introduction.

string	Literal match—interpreted relative to the current case-sensitivity lexical context (default is case-sensitive)
(*string1* *string2* ...)	Set of chars, *e.g.*, `("abc" "XYZ")`. Interpreted relative to the current case-sensitivity lexical context.
`(*` *sre* `...)`	0 or more matches
`(+` *sre* `...)`	1 or more matches
`(?` *sre* `...)`	0 or 1 matches
`(=` *n* *sre* `...)`	*n* matches
`(>=` *n* *sre* `...)`	*n* or more matches
`(**` *n* *m* *sre* `...)`	*n* to *m* matches
	N and *m* are Scheme expressions producing non-negative integers. *M* may also be `#f`, meaning "infinity."
`(|` *sre* `...)`	Choice (`or` is R5RS symbol;
`(or` *sre* `...)`	`|` is not specified by R5RS.)
`(:` *sre* `...)`	Sequence (`seq` is legal
`(seq` *sre* `...)`	Common Lisp symbol)
`(submatch` *sre* `...)`	Numbered submatch
`(dsm` *pre* *post* *sre* `...)`	Deleted submatches
	Pre and *post* are numerals.
`(uncase` *sre* `...)`	Case-folded match
`(w/case` *sre* `...)`	Introduce a lexical case-sensitivity
`(w/nocase` *sre* `...)`	context.
`,@`*exp*	Dynamically computed regexp
`,`*exp*	Same as ,@*exp*, but no submatch info
	Exp must produce a character, string, char-set, or regexp.
`bos eos`	Beginning/end of string
`bol eol`	Beginning/end of line
`bow eow`	Beginning/end of word

Figure 6.1: SRE syntax summary (part 1)

`(word` *sre* `...)`	(: bow *sre* … eow)
`(word+` *cset-sre* `...)`	`(word (+ (& (| alphanumeric "_")` `(|` *cset-sre* `...))))`
`word`	`(word+ any)`
`(posix-string` *string*`)`	Escape for Posix string notation
char	Singleton char set
class-name	alphanumeric, whitespace, *etc.*

These two forms are interpreted subject to the lexical case-sensitivity context.

`(~` *cset-sre* `...)`	Complement-of-union (`[^...]`)
`(-` *cset-sre* `...)`	Difference
`(&` *cset-sre* `...)`	Intersection
`(/` *range-spec* `...)`	Character range—interpreted subject to the lexical case-sensitivy context

Figure 6.2: SRE syntax summary (part 2)

```
class-name  ::= any
              |  nonl
              |  lower-case     | lower
              |  upper-case     | upper
              |  alphabetic     | alpha
              |  numeric        | digit | num
              |  alphanumeric   | alnum
              |  punctuation    | punct
              |  graphic        | graph
              |  whitespace     | space | white
              |  printing       | print
              |  control        | cntrl
              |  hex-digit      | xdigit | hex
              |  ascii
```

range-spec `::=` *string* `|` *char*
The chars are taken in pairs to form inclusive ranges.

Figure 6.3: SRE character-class names and range specs.

```
<cset-sre> ::= (~ <cset-sre> ...)    Set complement-of-union
             | (- <cset-sre> ...)    Set difference
             | (& <cset-sre> ...)    Intersection
             | (| <cset-sre> ...)    Set union
             | (/ <range-spec> ...)  Range

             | (<string>)            Constant set
             | <char>                Singleton constant set
             | <string>              For 1-char string "c"

             | <class-name>          Constant set

             | ,<exp>                <exp> evals to a char-set,
             | ,@<exp>               char, single-char string,
                                     or re-char-set regexp.

             | (uncase <cset-sre>)   Case-folding
             | (w/case <cset-sre>)
             | (w/nocase <cset-sre>)
```

Figure 6.4: applied to SRE's that specify character sets. These are the "type-checking" rules for character-set SRE's.

6.2 Examples

```
(- alpha ("aeiouAEIOU"))                ; Various forms of
(- alpha ("aeiou") ("AEIOU"))           ; non-vowel letter
(w/nocase (- alpha ("aeiou")))
(- (/"azAZ") ("aeiouAEIOU"))
(w/nocase (- (/"az") ("aeiou")))

;;; Upper-case letter, lower-case vowel, or digit
(| upper ("aeiou") digit)
(| (/"AZ09") ("aeiou"))

;;; Not an SRE, but Scheme code containing some embedded SREs.
(let* ((ws (rx (+ whitespace)))                 ; Seq of whitespace
       (date (rx (: (| "Jan" "Feb" "Mar" ...)   ; A month/day date.
                    ,ws
                    (| ("123456789")            ; 1-9
                       (: ("12") digit)         ; 10-29
                       "30" "31")))))           ; 30-31

  ;; Now we can use DATE several times:
  (rx ... ,date ... (* ... ,date ...)
      ... .... ,date))

;;; More Scheme code
(define (csl re)                ; A comma-separated list of RE's is
  (rx (| ""                     ; either zero of them (empty string), or
         (: ,re                 ; one RE, followed by
            (* ", " ,re)))))    ; Zero or more comma-space-RE matches.

(csl (rx (| "John" "Paul" "George" "Ringo")))
```

6.3 A short tutorial

S-expression regexps are called "SRE"s. Keep in mind that they are *not* Scheme expressions; they are another, separate notation that is expressed using the underlying framework of s-expression list structure: lists, symbols, *etc.* SRE's can be *embedded* inside of Scheme expressions using special forms that extend Scheme's syntax (such as the `rx` macro); there are places in the SRE grammar where one may place a Scheme expression. In these ways, SRE's and Scheme expressions can be intertwined. But this isn't fundamental; SRE's may be used in a completely Scheme-independent context. By simply restricting the notation to eliminate two special Scheme-embedding forms, they can be a completely independent notation.

Constant strings The simplest SRE is a string, denoting a constant regexp. For example, the SRE

```
"Spot"
```

matches only the string ≪capital-S, little-p, little-o, little-t≫. There is no interpretation of the characters in the string at all—the SRE

```
".*["
```

matches the string ≪period, asterisk, open-bracket≫.

Simple character sets To specify a set of characters, write a list whose single element is a string containing the set's elements. So the SRE

```
("aeiou")
```

only matches a vowel. One way to think of this, notationally, is that the set brackets are `("` and `")`.

Wild card Another simple SRE is the symbol `any`, which matches any single character—including newline and ASCII nul.

Sequences We can form sequences of SRE's with the SRE `(:` *sre* `...)`. So the SRE

```
(: "x" any "z")
```

matches any three-character string starting with "x" and ending with "z". As we'll see shortly, many SRE forms have bodies that are implicit sequences of other SRE's, analogous to the manner in which the body of a Scheme `lambda` or `let` expression is an implicit `begin` sequence. The regexp `(seq` *sre* `...)` is completely equivalent to `(:` *sre* `...)`; it's included in order to have a syntax that doesn't require `:` to be a legal symbol [1]

[1] That is, for use within s-expression syntax frameworks that, unlike R5RS, don't allow for `:` as a legal symbol. A Common Lisp embedding of SREs, for example, would need to use **`seq`** instead of `:`.

6.4 Choices

The SRE `(| sre ...)` is a regexp that matches anything any of the *sre* regexps match. So the regular expression

```
(| "sasha" "Pete")
```

matches either the string "sasha" or the string "Pete". The regexp

```
(| ("aeiou") ("0123456789"))
```

is the same as

```
("aeiou0123456789")
```

The regexp `(or sre ...)` is completely equivalent to `(| sre ...)`; it's included in order to have a syntax that doesn't require `|` to be a legal symbol.

Repetition There are several SRE forms that match multiple occurences of a regular expression. For example, the SRE `(* sre ...)` matches zero or more occurences of the sequence `(: sre ...)`. Here is the complete list of SRE repetition forms:

SRE	means	at least	no more than
`(*` *sre* `...)`	zero-or-more	0	infinity
`(+` *sre* `...)`	one-or-more	1	infinity
`(?` *sre* `...)`	zero-or-one	0	1
`(=` *from sre* `...)`	exactly-n	*from*	*from*
`(>=` *from sre* `...)`	n-or-more	*from*	infinity
`(**` *from to sre* `...)`	n-to-m	*from*	*to*

A *from* field is a Scheme expression that produces an integer. A *to* field is a Scheme expression that produces either an integer, or false, meaning infinity.

While it is illegal for the *from* or *to* fields to be negative, it *is* allowed for *from* to be greater than *to* in a `**` form—this simply produces a regexp that will never match anything.

As an example, we can describe the names of car/cdr access functions ("car", "cdr", "cadr", "cdar", "caar" , "cddr", "caaadr", *etc.*) with either of the SREs

```
(: "c" (+ (| "a" "d")) "r")
(: "c" (+ ("ad")) "r")
```

We can limit the a/d chains to 4 characters or less with the SRE

```
(: "c" (** 1 4 ("ad")) "r")
```

Some boundary cases:

```
(** 5 2 "foo")        ; Will never match
(** 0 0 "foo")        ; Matches the empty string
```

Character classes There is a special set of SRE's that form "character classes"—basically, a regexp that matches one character from some specified set of characters. There are operators to take the intersection, union, complement, and difference of character classes to produce a new character class. (Except for union, these capabilities are not provided for general regexps as they are computationally intractable in the general case.)

A single character is the simplest character class: `#\x` is a character class that matches only the character "x". A string that has only one letter is also a character class: `"x"` is the same SRE as `#\x`.

The character-set notation (*string*) we've seen is a primitive character class, as is the wildcard `any`. When arguments to the choice operator, `|`, are all character classes, then the choice form is itself a character-class. So these SREs are all character-classes:

```
("aeiou")
(| #\a #\e #\i #\o #\u)
(| ("aeiou") ("1234567890"))
```

However, these SRE's are *not* character-classes:

```
"aeiou"
(| "foo" #\x)
```

The `(~` *cset-sre* `...)` char class matches one character not in the specified classes:

```
(~ ("0248") ("1359"))
```

matches any character that is not a digit.

More compactly, we can use the `/` operator to specify character sets by giving the endpoints of contiguous ranges, where the endpoints are specified by a sequence of strings and characters. For example, any of these char classes

```
(/ #\A #\Z  #\a #\z  #\0 #\9)
(/ "AZ" #\a #\z "09")
(/ "AZ" #\a "z09")
(/"AZaz09")
```

matches a letter or a digit. The range endpoints are taken in pairs to form inclusive ranges of characters. Note that the exact set of characters included in a range is dependent on the underlying implementation's character type, so ranges may not be portable across different implementations.

There is a wide selection of predefined, named character classes that may be used. One such SRE is the wildcard `any`. `nonl` is a character class matching anything but newline; it is equivalent to

```
(~ #\newline)
```

and is useful as a wildcard in line-oriented matching.

There are also predefined named char classes for the standard Posix and Gnu character classes:

scsh name	Posix/ctype	Alternate name	Comment
`lower-case`	`lower`		
`upper-case`	`upper`		
`alphabetic`	`alpha`		
`numeric`	`digit`	`num`	
`alphanumeric`	`alnum`	`alphanum`	
`punctuation`	`punct`		
`graphic`	`graph`		
`blank`	(Gnu extension)		
`whitespace`	`space`	`white`	"`space`" is deprecated.
`printing`	`print`		
`control`	`cntrl`		
`hex-digit`	`xdigit`	`hex`	
`ascii`	(Gnu extension)		

See the scsh character-set documentation or the Posix isalpha(3) man page for the exact definitions of these sets.

You can use either the long scsh name or the shorter Posix and alternate names to refer to these char classes. The standard Posix name "`space`" is provided, but deprecated, since it is ambiguous. It means "whitespace," the set of whitespace characters, not the singleton set of the `#\space` character. If you want a short name for the set of whitespace characters, use the char-class name "white" instead.

Char classes may be intersected with the operator `(&` *cset-sre* `...)`, and set-difference can be performed with `(-` *cset-sre* `...)`. These operators are particularly useful when you want to specify a set by negation *with respect to a limited universe.* For example, the set of all non-vowel letters is

```
(- alpha ("aeiou") ("AEIOU"))
```

whereas writing a simple complement

```
(~ ("aeiouAEIOU"))
```

gives a char class that will match any non-vowel—including punctuation, digits, white space, control characters, and ASCII nul.

We can *compute* a char class by writing the SRE

,*cset-exp*

where *cset-exp* is a Scheme expression producing a value that can be coerced to a character set: a character set, character, one-character string, or char-class regexp value. This regexp matches one character from the set.

The char-class SRE `,@`*cset-exp* is entirely equivalent to `,`*cset-exp* when *cset-exp* produces a character set (but see below for the more general non-char-class context, where there *is* a distinction between `,`*exp* and `,@`*exp*.

As an example of character-class SREs, an SRE that matches a lower-case vowel, upper-case letter, or digit is

```
(| ("aeiou") (/"AZ09"))
```

or, equivalently

```
(| ("aeiou") upper-case numeric)
```

Boundary cases: the empty-complement char class

```
(~)
```

matches any character; it is equivalent to `any`. The empty-union char class

```
(|)
```

never matches at all. This is rarely useful for human-written regexps, but may be of occasional utility in machine-generated regexps, perhaps produced by macros.

The rules for determining if an SRE is a simple, char-class SRE or a more complex SRE form a little "type system" for SRE's. See the summary section preceding this one for a complete listing of these rules.

Case sensitivity There are three forms that control case sensitivity:

```
(uncase   sre ...)
(w/case   sre ...)
(w/nocase sre ...)
```

`uncase` is a regexp operator producing a regexp that matches any case permutation of any string that matches `(: sre ...)`. For example, the regexp

```
(uncase "foo")
```

matches the strings "foo", "foO", "fOo", "fOO", "Foo", ...

Expressions in SRE notation are interpreted in a lexical case-sensitivy context. The forms `w/case` and `w/nocase` are the scoping operators for this context, which controls how constant strings and char-class forms are interpreted in their bodies. So, for example, the regexp

```
(w/nocase "abc"
          (* "FOO" (w/case "Bar"))
          ("aeiou"))
```

defines a case-insensitive match for all of its elements except for the sub-element "Bar", which must match exactly capital-B, little-a, little-r. The default, the outermost, top-level context is case sensitive.

The lexical case-sensitivity context affects the interpretation of

- constant strings, such as `"foo"`,
- chars, such as `#\x`,
- char sets, such as `("abc")`, and
- ranges, such as `(/"az")` that appear within that context. It does not affect dynamically computed regexps—ones that are introduced by *,exp* and *,@exp* forms. It does not affect named char-classes—presumably, if you wrote `lower`, you didn't mean `alpha`.

  `uncase` is *not* the same as `w/nocase`. To point up one distinction, consider the two regexps

  ```
  (uncase   (~ "a"))
  (w/nocase (~ "a"))
  ```

The regexp `(~ "a")` matches any character except "a," which means it *does* match "A." Now, `(uncase` *re*`)` matches any case-permutation of a string that *re* matches. `(~ "a")` matches "A," so `(uncase (~ "a"))` matches "A" and "a"—and, for that matter, every other character. So `(uncase (~ "a"))` is equivalent to `any`.

In contrast, `(w/nocase (~ "a"))` establishes a case-insensitive lexical context in which the `"a"` is interpreted, making the SRE equivalent to `(~ ("aA"))`.

Dynamic regexps SRE notation allows you to compute parts of a regular expressions at run time. The SRE

`,`*exp*

is a regexp whose body *exp* is a Scheme expression producing a string, character, char-set, or regexp as its value. Strings and characters are converted into constant regexps; char-sets are converted into char-class regexps; and regexp values are substituted in place. So we can write regexps like this

```
(: "feeding the "
   ,(if (> n 1) "geese" "goose"))
```

This is how you can drop computed strings, such as someone's name, or the decimal numeral for a computed number, into a complex regexp.

If we have a large, complex regular expression that is used multiple times in some other, containing regular expression, we can name it, using the binding forms of the embedding language (*e.g.*, Scheme), and refer to it by name in the containing expression. For example, consider the Scheme expression

```
(let* ((ws (rx (+ whitespace)))  ; Seq of whitespace
       ;; Something like "Mar 14"
       (date (rx (: (| "Jan" "Feb" "Mar" ...)
                    ,ws
                    (| ("123456789")      ; 1-9
                       (: ("12") digit)   ; 10-29
                       "30"               ; 30
                       "31")))))          ; 31
  ;; Now we can use DATE several times:
  (rx ... ,date ... (* ... ,date ...)
      ... ,date ...))
```

where the `(rx` *sre* `...)` macro is the Scheme special form that produces a Scheme regexp value given a body in SRE notation.

As we saw in the char-class section, if a dynamic regexp is used in a char-class context (*e.g.*, as an argument to a `~` operation), the expression must be coercable not merely to a general regexp, but to a character sre—so it must be either a singleton string, a character, a scsh char set, or a char-class regexp.

We can also define and use functions on regexps in the host language. For example, consider the following Scheme expressions, containing embedded SRE's (inside the `rx` macro expressions) which in term contain embedded Scheme expressions computing dynamic regexps:

```
(define (csl re)
  ;; A comma-separated list of RE's is either
  (rx (| ""                 ; zero of them (empty string),
         (: ,re             ; or RE followed by
            (* ", " ,re))))); zero or more comma-space-RE matches.

(rx ... ,date ...
    ,(csl (rx (| "John" "Paul" "George" "Ringo")))
    ...
    ,(csl date)
    ...)
```

We leave the extension of `csl` to allow for an optional "and" between the last two matches as an exercise for the interested reader (*e.g.*, to match "John, Paul, George and Ringo").

Note, in passing, one of the nice features of SRE notation: they can be commented, and indented in a fashion to show the lexical extent of the subexpressions.

When we embed a computed regexp inside another regular expression with the *,exp* form, we must specify how to account for the submatches that may be in the computed part. For example, suppose we have the regexp

```
(rx (submatch (* "foo"))
    (submatch (? "bar"))
    ,(f x)
    (submatch "baz"))
```

It's clear that the submatch for the `(* "foo")` part of the regexp is submatch #1, and the `(? "bar")` part is submatch #2. But what number submatch is the `"baz"` submatch? It's not clear. Suppose the Scheme expression `(f x)` produces a regular expression that itself has 3 subforms. Are these counted (making the `"baz"` submatch #6), or not counted (making the `"bar"` submatch #3)?

SRE notation provides for both possibilities. The SRE

,exp

does *not* contribute its submatches to its containing regexp; it has zero submatches. So one can reliably assign submatch indices to forms appearing after a `,`*exp* form in a regexp.

On the other hand, the SRE

`,@`*exp*

"splices" its resulting regexp into place, *exposing* its submatches to the containing regexp. This is useful if the computed regexp is defined to produce a certain number of submatches—if that is part of *exp*'s "contract."

String, line, and word units The regexps `bos` and `eos` match the empty string at the beginning and end of the string, respectively.

The regexps `bol` and `eol` match the empty string at the beginning and end of a line, respectively. A line begins at the beginning of the string, and just after every newline character. A line ends at the end of the string, and just before every newline character. The char class `nonl` matches any character except newline, and is useful in conjunction with line-based pattern matching.

The regexps `bow` and `eow` match the empty string at the beginning and end of a word, respectively. A word is a contiguous sequence of characters that are either alphanumeric or the underscore character.

The regexp `(word` *sre* `...)` surrounds the sequence `(:` *sre* `...)`with bow/eow delimiters. It is equivalent to

```
(: bow sre ... eow)
```

The regexp `(word+` *cset-sre* `...)` matches a word whose body is one or more word characters matched by the char-set sre *cset-sre*. It is equivalent to

```
(word (+ (& (| alphanumeric "_")
            (| cset-sre ...))))
```

For example, a word not containing x, y, or z is

```
(word+ (~ ("xyz")))
```

The regexp `word` matches one word; it is equivalent to

```
(word+ any)
```

{Note `bol` and `eol` are not supported by scsh's current regexp search engine, which is Spencer's Posix matcher. This is the only element of the notation that is not supported by the current scsh reference implementation.}

Posix string notation The SRE `(posix-string` *string*`)`, where *string* is a string literal (*not* a general Scheme expression), allows one to use Posix string notation for a regexp. It's intended as backwards compatibility and is deprecated. For example, `(posix-string "[aeiou]+|x*|y{3,5}")` matches a string of vowels, a possibly empty string of x's, or three to five y's.

Note that parentheses are used ambiguously in Posix notation—both for grouping and submatch marking. The `(posix-string` *string*`)` form makes the conservative assumption: all parentheses introduce submatches.

Deleted submatches Deleted submatches, or "DSM's," are a subtle feature that are never required in expressions written by humans. They can be introduced by the simplifier when reducing regular expressions to simpler equivalents, and are included in the syntax to give it expressibility spanning the full regexp ADT. They may appear when unparsing simplified regular expressions that have been run through the simplifier; otherwise you are not likely to see them. Feel free to skip this section.

The regexp simplifier can sometimes eliminate entire sub-expressions from a regexp. For example, the regexp

```
(: "foo" (** 0 0 "apple") "bar")
```

can be simplified to

```
"foobar"
```

since `(** 0 0 "apple")` will always match the empty string. The regexp

```
(| "foo"
   (: "Richard" (|) "Nixon")
   "bar")
```

can be simplified to

```
(| "foo" "bar")
```

The empty choice `(|)` can't match anything, so the whole

```
(: "Richard" (|) "Nixon")
```

sequence can't match, and we can remove it from the choice.

However, if deleting part of a regular expression removes a submatch form, any following submatch forms will have their numbering changed, which would be an error. For example, if we simplify

```
(: (** 0 0 (submatch "apple"))
   (submatch "bar"))
```

to

```
(submatch "bar")
```

then the `"bar"` submatch changes from submatch #2 to submatch #1—so this is not a legal simplification.

When the simplifier deletes a sub-regexp that contains submatches, it introduces a special regexp form to account for the missing, deleted submatches, thus keeping the submatch accounting correct.

`(dsm` *pre post sre* `...)`

is a regexp that matches the sequence `(:` *sre* `...)`. *pre* and *post* are integer constants. The DSM form introduces *pre* deleted submatches before the body, and *post* deleted submatches after the body. If the body *(: sre ...)* itself has *body-sm* submatches, then the total number of submatches for the DSM form is

$$pre + body\text{-}sm + post.$$

These extra, deleted submatches are never assigned string indices in any match values produced when matching the regexp against a string.

As examples,

```
(| (: (submatch "Richard") (|) "Nixon")
   (submatch "bar"))
```

can be simplified to

```
(dsm 1 0 (submatch "bar"))
```

The regexp

```
(: (** 0 0 (submatch "apple"))
   (submatch "bar"))
```

can be simplified to

```
(dsm 1 0 (submatch "bar"))
```

6.4.1 Embedding regexps within Scheme programs

SRE's can be placed in a Scheme program using the `(rx` *sre* `...)` Scheme form, which evaluates to a Scheme regexp value.

Static and dynamic regexps

We separate SRE expressions into two classes: static and dynamic expressions. A *static* expression is one that has no run-time dependencies; it is a complete, self-contained description of a regular set. A *dynamic* expression is one that requires run-time computation to determine the particular regular set being described. There are two places where one can embed run-time computations in an SRE:

- The *from* or *to* repetition counts of `**`, `=`, and `>=` forms;
- `,`*exp* and `,@`*exp* forms.

A static SRE is one that does not contain any `,`*exp* or `,@`*exp* forms, and whose `**`, `=`, and `>=` forms all contain constant repetition counts.

Scsh's `rx` macro is able, at macro-expansion time, to completely parse, simplify and translate any static SRE into the equivalent Posix string which is used to drive the underlying C-based matching engine; there is no run-time overhead. Dynamic SRE's are partially simplified and then expanded into Scheme code that constructs the regexp at run-time.

6.5 Regexp functions

6.5.1 Obsolete, deprecated procedures

These two procedures are survivors from the previous, now-obsolete scsh regexp interface. Old code must open the `re-old-funs` package to access them. They should not be used in new code.

`(string-match` *posix-re-string string [start]*`)` $\longrightarrow$ *match or false* procedure
`(make-regexp` *posix-re-string*`)` $\longrightarrow$ *regexp* procedure

These are old functions included for backwards compatibility with previous releases. They are deprecated and will go away at some point in the future.

Note that the new release has no "regexp compiling" procedure at all—regexp values are compiled for the matching engine on-demand, and the necessary data structures are cached inside the ADT values.

6.5.2 Standard procedures and syntax

`(rx` *sre* …`)` $\longrightarrow$ *regexp* Syntax

This allows you to describe a regexp value with SRE notation.

(`regexp?` *x*) ⟶ *boolean* procedure

Returns true if the value is a regular expression.

(`regexp-search` *re string [start flags]*) ⟶ *match-data or false* procedure
(`regexp-search?` *re string [start flags]*) ⟶ *boolean* procedure

Search *string* starting at position *start*, looking for a match for regexp *re*. If a match is found, return a match structure describing the match, otherwise `#f`. *Start* defaults to 0.

Flags is the bitwise-or of `regexp/bos-not-bol` and `regexp/eos-not-eol`. `regexp/bos-not-bol` means the beginning of the string isn't a line-begin. `regexp/eos-not-eol` is analogous. {Note They're currently ignored because begining/end-of-line anchors aren't supported by the current implementation.}

Use `regexp-search?` when you don't need submatch information, as it has the potential to be *significantly* faster on submatch-containing regexps.

There is no longer a separate regexp "compilation" function; regexp values are compiled for the C engine on demand, and the resulting C structures are cached in the regexp structure after the first use.

(`match:start` *m [i]*) ⟶ *integer or false* procedure
(`match:end` *m [i]*) ⟶ *integer or false* procedure
(`match:substring` *m [i]*) ⟶ *string or false* procedure

`match:start` returns the start position of the submatch denoted by *match-number*. The whole regexp is 0; positive integers index submatches in the regexp, counting left-to-right. *Match-number* defaults to 0.

If the regular expression matches as a whole, but a particular subexpression does not match, then `match:start` returns `#f`.

`match:end` is analogous to `match:start`, returning the end position of the indexed submatch.

`match:substring` returns the substring matched regexp's submatch. If there was no match for the indexed submatch, it returns false.

(`regexp-substitute` *port-or-false match . items*) ⟶ *object* procedure

This procedure can be used to perform string substitutions based on regular-expression matches. The results of the substitution can be either output to a port or returned as a string.

The *match* argument is a regular-expression match structure that controls the substitution. If *port* is an output port, the *items* are written out to the port:

- If an item is a string, it is copied directly to the port.
- If an item is an integer, the corresponding submatch from *match* is written to the port.
- If an item is `'pre`, the prefix of the matched string (the text preceding the match) is written to the port.
- If an item is `'post`, the suffix of the matched string is written.

If *port* is `#f`, nothing is written, and a string is constructed and returned instead.

`(regexp-substitute/global` *port-or-false re str . items*`)` $\longrightarrow$ *object* procedure

This procedure is similar to `regexp-substitute`, but can be used to perform repeated match/substitute operations over a string. It has the following differences with `regexp-substitute`:

- It takes a regular expression and string to be matched as parameters, instead of a completed match structure.
- If the regular expression doesn't match the string, this procedure is the identity transform—it returns or outputs the string.
- If an item is `'post`, the procedure recurses on the suffix string (the text from *string* following the match). Including a `'post` in the list of items is how one gets multiple match/substitution operations.
- If an item is a procedure, it is applied to the match structure for a given match. The procedure returns a string to be used in the result.

The *regexp* parameter can be either a compiled regular expression or a string specifying a regular expression.

Some examples:

```
;;; Replace occurrences of "Cotton" with "Jin".
(regexp-substitute/global #f (rx "Cotton") s
                          'pre "Jin" 'post)

;;; mm/dd/yy -> dd/mm/yy date conversion.
(regexp-substitute/global #f (rx (submatch (+ digit)) "/" ; 1 = M
                                 (submatch (+ digit)) "/" ; 2 = D
                                 (submatch (+ digit)))    ; 3 = Y
                          s ; Source string
                          'pre 2 "/" 1 "/" 3 'post)

;;; "9/29/61" -> "Sep 29, 1961" date conversion.
(regexp-substitute/global #f (rx (submatch (+ digit)) "/" ; 1 = M
                                 (submatch (+ digit)) "/" ; 2 = D
                                 (submatch (+ digit)))    ; 3 = Y
                          s ; Source string
      'pre
      ;; Sleazy converter -- ignores "year 2000" issue,
      ;; and blows up if month is out of range.
      (lambda (m)
        (let ((mon (vector-ref '#("Jan" "Feb" "Mar" "Apr" "May" "Jun"
                                  "Jul" "Aug" "Sep" "Oct" "Nov" "Dec")
                               (- (string->number (match:substring m 1)) 1)))
              (day (match:substring m 2))
              (year (match:substring m 3)))
          (string-append mon " " day ", 19" year)))
      'post)

;;; Remove potentially offensive substrings from string S.
(define (kill-matches re s)
  (regexp-substitute/global #f s 'pre 'post))

(kill-matches (rx (| "Windows" "tcl" "Intel")) s)   ; Protect the children.
```

(`regexp-fold` *re kons knil s [finish start]*) $\longrightarrow$ *object* procedure

The following definition is a bit unwieldy, but the intuition is simple: this procedure uses the regexp *re* to divide up string *s* into non-matching/matching chunks, and then "folds" the procedure *kons* across this sequence of chunks. It is useful when you wish to operate on a string in sub-units defined by some regular expression, as are the related `regexp-fold-right` and `regexp-for-each` procedures.

Search from *start* (defaulting to 0) for a match to *re*; call this match *m*. Let *i* be the index of the end of the match (that is, (`match:end` *m* `0`)). Loop as follows:

(`regexp-fold` *re kons* (*kons start m knil*) *s finish i*)

If there is no match, return instead

(*finish start knil*)

Finish defaults to `(lambda (i knil) knil)`.

In other words, we divide up *s* into a sequence of non-matching/matching chunks:

$$NM_1\ M_1\ NM_1\ M_2\ \ldots\ NM_{k-1}\ M_{k-1}\ NM_k$$

where NM_1 is the initial part of *s* that isn't matched by the regexp *re*, M_1 is the first match, NM_2 is the following part of *s* that isn't matched, M_2 is the second match, and so forth—NM_k is the final non-matching chunk of *s*. We apply *kons* from left to right to build up a result, passing it one non-matching/matching chunk each time: on an application (*kons i m knil*), the non-matching chunk goes from *i* to `(match:begin` *m* `0)`, and the following matching chunk goes from `(match:begin` *m* `0)` to `(match:end` *m* `0)`. The last non-matching chunk NM_k is processed by *k*. So the computation we perform is

(*final Q* (*kons* j_k M_k ... (*kons* J_1 M_1 *knil*) ...))

where J_i is the index of the start of NM_i, M_i is a match value describing M_i, and Q is the index of the beginning of NM_k.

Hint: The `let-match` macro is frequently useful for operating on the match value *M* passed to the *kons* function.

`(regexp-fold-right` *re kons knil s [finish start]*`)` $\longrightarrow$ *object* procedure

The right-to-left variant of `regexp-fold`.

This procedure repeatedly matches regexp *re* across string *s*. This divides *s* up into a sequence of matching/non-matching chunks:

$$NM_1\ M_1\ NM_1\ M_2\ \ldots\ NM_{k-1}\ M_{k-1}\ NM_k$$

where NM_1 is the initial part of *s* that isn't matched by the regexp *re*, M_1 is the first match, NM_2 is the following part of *s* that isn't matched, M_2 is the second match, and so forth—NM_k is the final non-matching chunk of *s*. We apply *kons* from right to left to build up a result, passing it one non-matching/matching chunk each time:

(*final Q* (*kons* M_1 j_1 ... (*kons* M_k J_k *knil*) ...))

where MTCHi is a match value describing Mi, Ji is the index of the end of NMi (or, equivalently, the beginning of Mi+1), and Q is the index of the beginning of M1. In other words, KONS is passed a match, an index describing the following non-matching text, and the value produced by folding the following text. The FINAL function "polishes off" the fold operation by handling the initial chunk of non-matching text (NM0, above). FINISH defaults to (lambda (i knil) knil)

Example: To pick out all the matches to *re* in *s*, say

```
(regexp-fold-right re
                   (λ (m i lis)
                     (cons (match:substring m 0) lis))
                         '() s)
```

Hint: The `let-match` macro is frequently useful for operating on the match value *m* passed to the `kons` function.

(`regexp-for-each` *re proc s [start]*) ⟶ *undefined* procedure

Repeatedly match regexp *re* against string *s*. Apply *proc* to each match that is produced. Matches do not overlap.

Hint: The `let-match` macro is frequently useful for operating on the match value *m* passed to varproc.

(`let-match` *match-exp mvars body ...*) ⟶ *object* Syntax
(`if-match` *match-exp mvars on-match no-match*) ⟶ *object* Syntax

Mvars is a list of vars that is bound to the match and submatches of the string; `#F` is allowed as a don't-care element. For example,

```
(let-match (regexp-search date s) (whole-date month day year)
  ... body ...)
```

matches the regexp against string `s`, then evaluates the body of the `let-match` in a scope where `whole-date` is bound to the matched string, and `month`, `day` and `year` are bound to the first, second and third submatches.

`if-match` is similar, but if the match expression is false, then the *no-match* expression is evaluated; this would be an error in `let-match`.

(`match-cond` *clause ...*) ⟶ *object* Syntax

This macro allows one to conditionally attempt a sequence of pattern matches, interspersed with other, general conditional tests. There are four kinds of `match-cond` clause, one introducing a pattern match, and the other three simply being regular `cond`-style clauses, marked by the `test` and `else` keywords:

```
(match-cond (match-exp match-vars body ...) ; As in if-match
            (test exp body ...)             ; As in cond
            (test exp => proc)              ; As in cond
            (else body ...))                ; As in cond
```

(flush-submatches *re*) $\longrightarrow$ *re* procedure
(uncase *re*) $\longrightarrow$ *re* procedure
(simplify-regexp *re*) $\longrightarrow$ *re* procedure
(uncase-char-set *cset*) $\longrightarrow$ *re* procedure
(uncase-string *str*) $\longrightarrow$ *re* procedure

These functions map regexps and char sets to other regexps. `flush-submatches` returns a regexp which matches exactly what its argument matches, but contains no submatches.

`uncase` returns a regexp that matches any case-permutation of its argument regexp.

`simplify-regexp` applies the simplifier to its argument. This is done automatically when compiling regular expressions, so this is only useful for programmers that are directly examining the ADT value with lower-level accessors.

`uncase-char-set` maps a char set to a regular expression that matches any character from that set, regardless of case. Similarly, `uncase-string` returns a regexp that matches any case-permutation of the string. For example, `(uncase-string "Knight")` returns the same value that `(rx ("kK") ("nN") ("iI") ("gG") ("hH") ("tT"))` or `(rx (w/nocase "Knight"))`.

(sre->regexp *sre*) $\longrightarrow$ *re* procedure
(regexp->sre *re*) $\longrightarrow$ *sre* procedure

These are the SRE parser and unparser. That is, `sre->regexp` maps an SRE to a regexp value, and `regexp->sre` does the inverse. The latter function can be useful for printing out regexps in a readable format.

```
(sre->regexp '(: "Olin " (? "G. ") "Shivers")) ⟹  regexp
(define re (re-seq (re-string "Pete ")
                   (re-repeat 1 #f (re-string "Sz"))
                   (re-string "ilagyi")))
(regexp->sre (re-repeat 0 1 re))
    ⟹  '(? "Pete" (+ "Sz") "ilagyi")
```

(posix-string->regexp *string*) $\longrightarrow$ *re* procedure
(regexp->posix-string *re*) $\longrightarrow$ *string* procedure

These two functions are the Posix notation parser and unparser. That is, `posix-string->regexp` maps a Posix-notation regular expression, such as `"g(ee|oo)se"`, to a regexp value, and `regexp->posix-string` does the inverse.

You can use these tools to map between scsh regexps and Posix regexp strings, which can be useful if you want to do conversion between SRE's and Posix form. For example, you can write a particularly complex regexp in SRE form, or compute it using the ADT constructors, then convert to Posix form, print it out, cut and paste it into a C or emacs lisp program. Or you can import an old regexp from some other program, parse it into an ADT value, render it to an SRE, print it out, then cut and paste it into a scsh program.

Note:

- The string parser doesn't handle the exotica of character class names such as `[[:alnum:]]`; the current implementation was written in in three hours.
- The unparser produces Spencer-specific strings for bow/eow elements; otherwise, it's Posix all the way.

6.6 The regexp ADT

The following functions may be used to construct and examine scsh's regexp abstract data type. They are in the following Scheme 48 packages: re-adt-lib re-lib scsh

Each basic class of regexp has a predicate, a basic constructor, a "smart" consructor that performs limited "peephole" optimisation on its arguments, and a set of accessors. The `...:tsm` accessor returns the total number of submatches contained in the regular expression.

`(re-seq?` *x*`)` ⟶ *boolean*	Type predicate
`(make-re-seq` *re ...*`)` ⟶ *re*	Basic constructor
`(re-seq` *re ...*`)` ⟶ *re*	Smart constructor
`(re-seq:elts` *re*`)` ⟶ *re-list*	Accessor
`(re-seq:tsm` *re*`)` ⟶ *integer*	Accessor
`(re-choice?` *x*`)` ⟶ *boolean*	Type predicate
`(make-re-choice` *re-list*`)` ⟶ *re*	Basic constructor
`(re-choice` *re ...*`)` ⟶ *re*	Smart constructor
`(re-choice:elts` *re*`)` ⟶ *re-list*	Accessor
`(re-choice:tsm` *re*`)` ⟶ *integer*	Accessor
`(re-repeat?` *x*`)` ⟶ *boolean*	Type predicate
`(make-re-repeat` *from to body*`)` ⟶ *re*	Accessor
`(re-repeat:from` *re*`)` ⟶ *integer*	Accessor
`(re-repeat:to` *re*`)` ⟶ *integer*	Accessor
`(re-repeat:tsm` *re*`)` ⟶ *integer*	Accessor
`(re-submatch?` *x*`)` ⟶ *boolean*	Type predicate
`(make-re-submatch` *body [pre-dsm post-dsm]*`)` ⟶ *re*	Accessor

(`re-submatch:pre-dsm` *re*) ⟶ *integer* Accessor
(`re-submatch:post-dsm` *re*) ⟶ *integer* Accessor
(`re-submatch:tsm` *re*) ⟶ *integer* Accessor

(`re-string?` *x*) ⟶ *boolean* Type predicate
(`make-re-string` *chars*) ⟶ *re* Basic constructor
(`re-string` *chars*) ⟶ *re* Basic constructor
(`re-string:chars` *re*) ⟶ *string* Accessor

(`re-char-set?` *x*) ⟶ *boolean* Type predicate
(`make-re-char-set` *cset*) ⟶ *re* Basic constructor
(`re-char-set` *cset*) ⟶ *re* Basic constructor
(`re-char-set:cset` *re*) ⟶ *char-set* Accessor

(`re-dsm?` *x*) ⟶ *boolean* Type predicate
(`make-re-dsm` *body pre-dsm post-dsm*) ⟶ *re* Basic constructor
(`re-dsm` *body pre-dsm post-dsm*) ⟶ *re* Smart constructor
(`re-dsm:body` *re*) ⟶ *re* Accessor
(`re-dsm:pre-dsm` *re*) ⟶ *integer* Accessor
(`re-dsm:post-dsm` *re*) ⟶ *integer* Accessor
(`re-dsm:tsm` *re*) ⟶ *integer* Accessor

`re-bos` regexp
`re-eos` regexp
`re-bol` regexp
`re-eol` regexp
`re-bow` regexp
`re-eow` regexp

These variables are bound to the primitive anchor regexps.

(`re-bos?` *object*) ⟶ *boolean* procedure
(`re-eos?` *object*) ⟶ *boolean* procedure
(`re-bol?` *object*) ⟶ *boolean* procedure
(`re-eol?` *object*) ⟶ *boolean* procedure
(`re-bow?` *object*) ⟶ *boolean* procedure
(`re-eow?` *object*) ⟶ *boolean* procedure

These predicates recognise the associated primitive anchor regexp.

`re-trivial` regexp
(`re-trivial?` *re*) ⟶ *boolean* procedure

The variable `re-trivial` is bound to a regular expression that matches the empty string (corresponding to the SRE `""` or `(:)`); it is recognised by the associated predicate. Note that the predicate is only guaranteed to recognise this particular trivial regexp; other trivial regexps built using other constructors may or may not produce a true value.

`re-empty` regexp
`(re-empty? re)` ⟶ *boolean* procedure

The variable `re-empty` is bound to a regular expression that never matches (corresponding to the SRE `(|)`); it is recognised by the associated predicate. Note that the predicate is only guaranteed to recognise this particular empty regexp; other empty regexps built using other constructors may or may not produce a true value.

`re-any` regexp
`(re-any? re)` ⟶ *boolean* procedure

The variable `re-any` is bound to a regular expression that matches the any character (corresponding to the SRE `any`); it is recognised by the associated predicate. Note that the predicate is only guaranteed to recognise this particular any-character regexp value; other any-character regexps built using other constructors may or may not produce a true value.

`re-nonl` regexp
`re-word` regexp

The variable `re-any` is bound to a regular expression that matches the any non-newline character (corresponding to the SRE `(~ #\newline)`).

Similarly, `re-word` is bound to a regular expression that matches any word (corresponding to the SRE `word`).

`(regexp? object)` ⟶ *boolean* procedure

Is the object a regexp?

`(re-tsm re)` ⟶ *integer* procedure

Return the total number of submatches contained in the regexp.

`(clean-up-cres)` ⟶ *undefined* procedure

The current scsh implementation should call this function periodically to release C-heap storage associated with compiled regexps. Hopefully, this procedure will be removed at a later date.

6.7 Syntax-hacking tools

The Scheme 48 package `rx-syntax-tools` exports several tools for macro writers that want to use SREs in their macros. In the functions defined below, *compare* and *rename* parameters are as passed to Clinger-Rees explicit-renaming low-level macros.

`(if-sre-form form conseq-form alt-form)` ⟶ *form* Syntax

If *form* is a legal SRE, this is equivalent to the expression *conseq-form*, otherwise it expands to *alt-form*.

This is useful for high-level macro authors who want to write a macro where one field in the macro can be an SRE or possibly something else. *E.g.*, we might have a conditional form wherein if the test part of one arm is an SRE, it expands to a regexp match on some implied value, otherwise the form is evaluated as a boolean Scheme expression. For example, a conditional macro might expand into code containing the following form, which in turn would have one of two possible expansions:

```
(if-sre-form test-exp                  ; If TEST-EXP is SRE,
  (regexp-search? (rx test-exp) line) ; match it w/the line,
  test-exp)                            ; otw it's a text exp.
```

(`sre-form?` *form rename compare*) $\longrightarrow$ *boolean* procedure

This procedure is for low-level macros doing things equivalent to `if-sre-form`. It returns true if the form is a legal SRE.

Note that neither `sre-form` nor `if-sre-form` does a deep recursion over the form in the case where the form is a list. They simply check the car of the form for one of the legal SRE keywords.

(`parse-sre` *sre-form compare rename*) $\longrightarrow$ *re* procedure
(`parse-sres` *sre-forms compare rename*) $\longrightarrow$ *re* procedure

Parse `sre-form` into an ADT. Note that if the SRE is dynamic—contains `,`*exp* or `,@`*exp* forms, or has repeat operators whose from/to counts are not constants—then the returned ADT will have *Scheme expressions* in the corresponding slots of the regexp records instead of the corresponding integer, char-set, or regexp. In other words, we use the ADT as its own AST. It's called a "hack."

`parse-sres` parses a list of SRE forms that comprise an implicit sequence.

(`regexp->scheme` *re rename*) $\longrightarrow$ *Scheme-expression* procedure

Returns a Scheme expression that will construct the regexp *re* using ADT constructors such as `make-re-sequence`, `make-re-repeat`, and so forth.

If the regexp is static, it will be simplified and pre-translated to a Posix string as well, which will be part of the constructed regexp value.

(`static-regexp?` *re*) $\longrightarrow$ *boolean* procedure

Is the regexp a static one?

Chapter 7

Reading delimited strings

Scsh provides a set of procedures that read delimited strings from input ports. There are procedures to read a single line of text (terminated by a newline character), a single paragraph (terminated by a blank line), and general delimited strings (terminated by a character belonging to an arbitrary character set).

These procedures can be applied to any Scheme input port. However, the scsh virtual machine has native-code support for performing delimited reads on Unix ports, and these input operations should be particularly fast—much faster than doing the equivalent character-at-a-time operation from Scheme code.

All of the delimited input operations described below take a `handle-delim` parameter, which determines what the procedure does with the terminating delimiter character. There are four possible choices for a `handle-delim` parameter:

`handle-delim`	Meaning
`'trim`	Ignore delimiter character.
`'peek`	Leave delimiter character in input stream.
`'concat`	Append delimiter character to returned value.
`'split`	Return delimiter as second value.

The first case, `'trim`, is the standard default for all the routines described in this section. The last three cases allow the programmer to distinguish between strings that are terminated by a delimiter character, and strings that are terminated by an end-of-file.

(`read-line` *[port handle-newline]*) $\longrightarrow$ *string or eof-object* procedure

Reads and returns one line of text; on eof, returns the eof object. A line is terminated by newline or eof.

handle-newline determines what `read-line` does with the newline or EOF that terminates the line; it takes the general set of values described for the general `handle-delim` case above, and defaults to `'trim` (discard the newline). Using this argument allows one to tell whether or not the last line of input in a file is newline terminated.

(`read-paragraph` *[port handle-delim]*) $\longrightarrow$ *string or eof* procedure

This procedure skips blank lines, then reads text from a port until a blank line or eof is found. A "blank line" is a (possibly empty) line composed only of white space. The *handle-delim* parameter determines how the terminating blank line is handled. It is described above, and defaults to `'trim`. The `'peek` option is not available.

The following procedures read in strings from ports delimited by characters belonging to a specific set. See section 5.3 for information on character set manipulation.

(`read-delimited` *char-set [port handle-delim]*) $\longrightarrow$ *string or eof* procedure

Read until we encounter one of the chars in *char-set* or eof. The *handle-delim* parameter determines how the terminating character is handled. It is described above, and defaults to `'trim`.

The *char-set* argument may be a charset, a string, a character, or a character predicate; it is coerced to a charset.

(`read-delimited!` *char-set buf [port handle-delim start end]*) $\longrightarrow$ *nchars or eof or #f* procedure

A side-effecting variant of `read-delimited`.

The data is written into the string *buf* at the indices in the half-open interval $[start, end)$; the default interval is the whole string: $start = 0$ and $end =$ (`string-length` *buf*). The values of *start* and *end* must specify a well-defined interval in *str*, *i.e.*, $0 \leq start \leq end \leq$ (`string-length` *buf*).

It returns *nbytes*, the number of bytes read. If the buffer filled up without a delimiter character being found, `#f` is returned. If the port is at eof when the read starts, the eof object is returned.

If an integer is returned (*i.e.*, the read is successfully terminated by reading a delimiter character), then the *handle-delim* parameter determines how the terminating character is handled. It is described above, and defaults to `'trim`.

(`%read-delimited!` *char-set buf gobble? [port start end]*) $\longrightarrow$ *[char-or-eof-or-#f integer]* procedure

This low-level delimited reader uses an alternate interface. It returns two values: *terminator* and *num-read*.

terminator A value describing why the read was terminated:

Character or eof-object	$\Rightarrow$	Read terminated by this value.
`#f`	$\Rightarrow$	Filled buffer without finding a delimiter.

num-read Number of characters read into *buf*.

If the read is successfully terminated by reading a delimiter character, then the *gobble?* parameter determines what to do with the terminating character. If true, the character is removed from the input stream; if false, the character is left in the input stream where a subsequent read operation will retrieve it. In either case, the character is also the first value returned by the procedure call.

`(skip-char-set` *skip-chars [port]*`)` $\longrightarrow$ *integer* procedure

Skip characters occurring in the set *skip-chars*; return the number of characters skipped. The *skip-chars* argument may be a charset, a string, a character, or a character predicate; it is coerced to a charset.

Chapter 8

Awk, record I/O, and field parsing

Unix programs frequently process streams of records, where each record is delimited by a newline, and records are broken into fields with other delimiters (for example, the colon character in `/etc/passwd`). Scsh has procedures that allow the programmer to easily do this kind of processing. Scsh's field parsers can also be used to parse other kinds of delimited strings, such as colon-separated `$PATH` lists. These routines can be used with scsh's `awk` loop construct to conveniently perform pattern-directed computation over streams of records.

8.1 Record I/O and field parsing

The procedures in this section are used to read records from I/O streams and parse them into fields. A record is defined as text terminated by some delimiter (usually a newline). A record can be split into fields by using regular expressions in one of several ways: to *match* fields, to *separate* fields, or to *terminate* fields. The field parsers can be applied to arbitrary strings (one common use is splitting environment variables such as `$PATH` at colons into its component elements).

The general delimited-input procedures described in chapter 7 are also useful for reading simple records, such as single lines, paragraphs of text, or strings terminated by specific characters.

8.1.1 Reading records

`(record-reader` *[delims elide-delims? handle-delim]*`)` $\longrightarrow$ *procedure* procedure

Returns a procedure that reads records from a port. The procedure is invoked as follows:

> (*reader* *[port]*) $\longrightarrow$ *string or eof*

A record is a sequence of characters terminated by one of the characters in *delims* or eof. If *elide-delims?* is true, then a contiguous sequence of delimiter chars are taken as a single record delimiter. If *elide-delims?* is false, then a delimiter char coming immediately after a delimiter char produces an empty-string record. The reader consumes the delimiting char(s) before returning from a read.

The *delims* set defaults to the set {newline}. It may be a charset, string, character, or character predicate, and is coerced to a charset. The *elide-delims?* flag defaults to `#f`.

The *handle-delim* argument controls what is done with the record's terminating delimiter.

`'trim`	Delimiters are trimmed. (The default)
`'split`	Reader returns delimiter string as a second argument. If record is terminated by EOF, then the eof object is returned as this second argument.
`'concat`	The record and its delimiter are returned as a single string.

The reader procedure returned takes one optional argument, the port from which to read, which defaults to the current input port. It returns a string or eof.

8.1.2 Parsing fields

`(field-splitter` *[field num-fields]*`)` $\longrightarrow$ *procedure* procedure
`(infix-splitter` *[delim num-fields handle-delim]*`)` $\longrightarrow$ *procedure* procedure
`(suffix-splitter` *[delim num-fields handle-delim]*`)` $\longrightarrow$ *procedure* procedure
`(sloppy-suffix-splitter` *[delim num-fields handle-delim]*`)` $\longrightarrow$ *procedure* procedure

These functions return a parser function that can be used as follows:

> (*parser* *string* *[start]*) $\longrightarrow$ *string-list*

The returned parsers split strings into fields defined by regular expressions. You can parse by specifying a pattern that *separates* fields, a pattern that *terminates* fields, or a pattern that *matches* fields:

Procedure	Pattern
`field-splitter`	matches fields
`infix-splitter`	separates fields
`suffix-splitter`	terminates fields
`sloppy-suffix-splitter`	terminates fields

These parser generators are controlled by a range of options, so that you can precisely specify what kind of parsing you want. However, these options default to reasonable values for general use.

Defaults:

delim	=	`(rx (| (+ white) eos))`	(suffix delimiter: white space or eos)
		`(rx (+ white))`	(infix delimiter: white space)
field	=	`(rx (+ (~ white)))`	(non-white-space)
num-fields	=	`#f`	(as many fields as possible)
handle-delim	=	`'trim`	(discard delimiter chars)

. . . which means: break the string at white space, discarding the white space, and parse as many fields as possible.

The *delim* parameter is a regular expression matching the text that occurs between fields. See chapter 6 for information on regular expressions, and the `rx` form used to specify them. In the separator case, it defaults to a pattern matching white space; in the terminator case, it defaults to white space or end-of-string.

The *field* parameter is a regular expression used to match fields. It defaults to non-white-space.

The *delim* patterns may also be given as a string, character, or char-set, which are coerced to regular expressions. So the following expressions are all equivalent, each producing a function that splits strings apart at colons:

```
(infix-splitter (rx ":"))
(infix-splitter ":")
(infix-splitter #\:)
(infix-splitter (char-set #\:))
```

The boolean *handle-delim* determines what to do with delimiters.

`'trim`	Delimiters are thrown away after parsing. (default)
`'concat`	Delimiters are appended to the field preceding them.
`'split`	Delimiters are returned as separate elements in the field list.

The *num-fields* argument used to create the parser specifies how many fields to parse. If `#f` (the default), the procedure parses them all. If a positive integer n, exactly that many fields are parsed; it is an error if there are more or fewer than n fields in the record. If *num-fields* is a negative integer or zero, then $|n|$ fields are parsed, and the remainder of the string is returned in the last element of the field list; it is an error if fewer than $|n|$ fields can be parsed.

The field parser produced is a procedure that can be employed as follows:

(*parse string [start]*) $\Longrightarrow$ *string-list*

The optional *start* argument (default 0) specifies where in the string to begin the parse. It is an error if *start* $>$ `(string-length` *string*`)`.

The parsers returned by the four parser generators implement different kinds of field parsing:

`field-splitter` The regular expression specifies the actual field.

`suffix-splitter` Delimiters are interpreted as element *terminators*. If vertical-bar is the the delimiter, then the string `""` is the empty record `()`, `"foo|"` produces a one-field record `("foo")`, and `"foo"` is an error.

The syntax of suffix-delimited records is:

<record>	::=	`""`	(Empty record)
	|	*<element> <delim> <record>*	

It is an error if a non-empty record does not end with a delimiter. To make the last delimiter optional, make sure the delimiter regexp matches the end-of-string (sre `eos`).

`infix-splitter` Delimiters are interpreted as element *separators*. If comma is the delimiter, then the string `"foo,"` produces a two-field record `("foo" "")`.

The syntax of infix-delimited records is:

<record>	::=	`""`	(Forced to be empty record)
	|	*<real-infix-record>*	
<real-infix-record>	::=	*<element> <delim> <real-infix-record>*	
	|	*<element>*	

Note that separator semantics doesn't really allow for empty records—the straightforward grammar (*i.e.*, *<real-infix-record>*) parses an empty string as a singleton list whose one field is the empty string, `("")`, not as the empty record `()`. This is unfortunate, since it means that infix string parsing doesn't make `string-append` and `append` isomorphic. For example,

```
((infix-splitter ":") (string-append x ":" y))
```

doesn't always equal

```
(append ((infix-splitter ":") x)
        ((infix-splitter ":") y))
```

Record	: suffix	:|$ suffix	: infix	non-: field
`""`	`()`	`()`	`()`	`()`
`":"`	`("")`	`("")`	`("" "")`	`()`
`"foo:"`	`("foo")`	`("foo")`	`("foo" "")`	`("foo")`
`":foo"`	*error*	`("" "foo")`	`("" "foo")`	`("foo")`
`"foo:bar"`	*error*	`("foo" "bar")`	`("foo" "bar")`	`("foo" "bar")`

Figure 8.1: Using different grammars to split records into fields.

It fails when x or y are the empty string. Terminator semantics *does* preserve a similar isomorphism.

However, separator semantics is frequently what other Unix software uses, so to parse their strings, we need to use it. For example, Unix `$PATH` lists have separator semantics. The path list `"/bin:"` is broken up into `("/bin" "")`, not `("/bin")`. Comma-separated lists should also be parsed this way.

`sloppy-suffix` The same as the `suffix` case, except that the parser will skip an initial delimiter string if the string begins with one instead of parsing an initial empty field. This can be used, for example, to field-split a sequence of English text at white-space boundaries, where the string may begin or end with white space, by using regex

```
(rx (| (+ white) eos))
```

(But you would be better off using `field-splitter` in this case.)

Figure 8.1 shows how the different parser grammars split apart the same strings. Having to choose between the different grammars requires you to decide what you want, but at least you can be precise about what you are parsing. Take fifteen seconds and think it out. Say what you mean; mean what you say.

(`join-strings` *string-list [delimiter grammar]*) $\longrightarrow$ *string* procedure

This procedure is a simple unparser—it pastes strings together using the delimiter string.

The *grammar* argument is one of the symbols `infix` (the default) or `suffix`; it determines whether the delimiter string is used as a separator or as a terminator.

The delimiter is the string used to delimit elements; it defaults to a single space `" "`.

Example:

```
(join-strings '("foo" "bar" "baz") ":")
    ⟹  "foo:bar:baz"
```

8.1.3 Field readers

(`field-reader` *[field-parser rec-reader]*) ⟶ *procedure* procedure

This utility returns a procedure that reads records with field structure from a port. The reader's interface is designed to make it useful in the `awk` loop macro (section 8.2). The reader is used as follows:

(*reader [port]*) ⟹ *[raw-record parsed-record]* `or` *[eof ()]*

When the reader is applied to an input port (default: the current input port), it reads a record using *rec-reader*. If this record isn't the eof object, it is parsed with *field-parser*. These two values—the record, and its parsed representation—are returned as multiple values from the reader.

When called at eof, the reader returns [eof-object `()`].

Although the record reader typically returns a string, and the field-parser typically takes a string argument, this is not required. The record reader can produce, and the field-parser consume, values of any type. However, the empty list returned as the parsed value on eof is hardwired into the field reader.

For example, if port `p` is open on `/etc/passwd`, then

```
((field-reader (infix-splitter ":" 7)) p)
```

returns two values:

```
"dalbertz:mx3Uaqq0:107:22:David Albertz:/users/dalbertz:/bin/csh"
("dalbertz" "mx3Uaqq0" "107" "22" "David Albertz" "/users/dalbertz"
           "/bin/csh")
```

The *field-parser* defaults to the value of `(field-splitter)`, a parser that picks out sequences of non-white-space strings.

The *rec-reader* defaults to `read-line`.

Figure 8.2 shows `field-reader` being used to read different kinds of Unix records.

8.1.4 Forward-progress guarantees and empty-string matches

A loop that pulls text off a string by repeatedly matching a regexp against that string can conceivably get stuck in an infinite loop if the regexp matches the empty string. For example, the SREs `bos`, `eos`, `(* any)`, and `(| "foo" (* (  "f")))` can all match the empty string.

```
;;; /etc/passwd reader
(field-reader (infix-splitter ":" 7))
    ; wandy:3xuncWdpKhR.:73:22:Wandy Saetan:/usr/wandy:/bin/csh

;;; Two ls -l output readers
(field-reader (infix-splitter (rx (+ white)) 8))
(field-reader (infix-splitter (rx (+ white)) -7))
    ; -rw-r--r--  1 shivers    22880 Sep 24 12:45 scsh.scm

;;; Internet hostname reader
(field-reader (field-splitter (rx (+ (~ ".")))))
    ; stat.sinica.edu.tw

;;; Internet IP address reader
(field-reader (field-splitter (rx (+ (~ "."))) 4))
    ; 18.24.0.241

;;; Line of integers
(let ((parser (field-splitter (rx (? ("+-")) (+ digit)))))
  (field-reader (λ (s) (map string->number (parser s))))
    ; 18 24 0 241

;;; Same as above.
(let ((reader (field-reader (field-splitter (rx (? ("+-"))
                                                 (+ digit))))))
  (λ maybe-port (map string->number (apply reader maybe-port))))
    ; Yale beat harvard 26 to 7.
```

Figure 8.2: Some examples of `field-reader`

The routines in this package that iterate through strings with regular expressions are careful to handle this empty-string case. If a regexp matches the empty string, the next search starts, not from the end of the match (which in the empty string case is also the beginning—that's the problem), but from the next character over. This is the correct behaviour. Regexps match the longest possible string at a given location, so if the regexp matched the empty string at location i, then it is guaranteed it could not have matched a longer pattern starting with character i. So we can safely begin our search for the next match at char $i + 1$.

With this provision, every iteration through the loop makes some forward progress, and the loop is guaranteed to terminate.

This has the effect you want with field parsing. For example, if you split a string with the empty pattern, you will explode the string into its individual characters:

```
((suffix-splitter (rx)) "foo") ⟹  ("" "f" "o" "o")
```

However, even though this boundary case is handled correctly, we don't recommend using it. Say what you mean—just use a field splitter:

```
((field-splitter (rx any)) "foo") ⟹  ("f" "o" "o")
```

Or, more efficiently,

```
((λ (s) (map string (string->list s))) "foo")
```

8.1.5 Reader limitations

Since all of the readers in this package require the ability to peek ahead one char in the input stream, they cannot be applied to raw integer file descriptors, only Scheme input ports. This is because Unix doesn't support peeking ahead into input streams.

8.2 Awk

Scsh provides a loop macro and a set of field parsers that can be used to perform text processing very similar to the Awk programming language. The basic functionality of Awk is factored in scsh into its component parts. The control structure is provided by the `awk` loop macro; the text I/O and parsers are provided by the field-reader subroutine library (section 8.1). This factoring allows the programmer to compose the basic loop structure with any parser or input mechanism at all. If the parsers provided by the field-reader package are insufficient, the programmer can write a custom parser in Scheme and use it with equal ease in the awk framework.

Awk-in-scheme is given by a loop macro called `awk`. It looks like this:

```
(awk <next-record> <record&field-vars>
     [<counter>] <state-var-decls>
  <clause1> ...)
```

The body of the loop is a series of clauses, each one representing a kind of condition/action pair. The loop repeatedly reads a record, and then executes each clause whose condition is satisfied by the record.

Here's an example that reads lines from port `p` and prints the line number and line of every line containing the string "`Church-Rosser`":

```
(awk (read-line) (ln) lineno ()
  ("Church-Rosser" (format #t "~d: ~s~%" lineno ln)))
```

This example has just one clause in the loop body, the one that tests for matches against the regular expression "`Church-Rosser`".

The *<next-record>* form is an expression that is evaluated each time through the loop to produce a record to process. This expression can return multiple values; these values are bound to the variables given in the *<record&field-vars>* list of variables. The first value returned is assumed to be the record; when it is the end-of-file object, the loop terminates.

For example, let's suppose we want to read items from `/etc/password`, and we use the `field-reader` procedure to define a record parser for `/etc/passwd` entries:

```
(define read-passwd (field-reader (infix-splitter ":" 7)))
```

binds `read-passwd` to a procedure that reads in a line of text when it is called, and splits the text at colons. It returns two values: the entire line read, and a seven-element list of the split-out fields. (See section 8.1 for more on `field-reader` and `infix-splitter`.)

So if the *<next-record>* form in an `awk` expression is `(read-passwd)`, then *<record&field-vars>* must be a list of two variables, *e.g.*,

```
(record field-vec)
```

since `read-passwd` returns two values.

Note that `awk` allows us to use *any* record reader we want in the loop, returning whatever number of values we like. These values don't have to be strings or string lists. The only requirement is that the record reader return the eof object as its first value when the loop should terminate.

The `awk` loop allows the programmer to have loop variables. These are declared and initialised by the *<state-var-decls>* form, a

```
((var init-exp) (var init-exp) ...)
```

list rather like the `let` form. Whenever a clause in the loop body executes, it evaluates to as many values as there are state variables, updating them.

The optional *<counter>* variable is an iteration counter. It is bound to 0 when the loop starts. The counter is incremented each time a non-eof record is read.

There are several kinds of loop clause. When evaluating the body of the loop, `awk` evaluates *all* the clauses sequentially. Unlike `cond`, it does not stop after the first clause is satisfied; it checks them all.

- `(`*test body*$_1$ *body*$_2$ `...)`
  If *test* is true, execute the body forms. The last body form is the value of the clause. The test and body forms are evaluated in the scope of the record and state variables.

  The *test* form can be one of:

integer:	The test is true for that iteration of the loop. The first iteration is #1.
sre:	A regular expression, in SRE notation (see chapter 6) can be used as a test. The test is successful if the pattern matches the record. In particular, note that any string is an SRE.
`(when` *expr*`)`:	The body of a `when` test is evaluated as a Scheme boolean expression in the inner scope of the `awk` form.
expr:	If the form is none of the above, it is treated as a Scheme expression— in practice, the `when` keyword is only needed in cases where SRE/Scheme expression ambiguity might occur.

- `(range` *start-test stop-test body*$_1$ `...)`
  `(:range` *start-test stop-test body*$_1$ `...)`
  `(range:` *start-test stop-test body*$_1$ `...)`
  `(:range:` *start-test stop-test body*$_1$ `...)`
  These clauses become activated when *start-test* is true; they stay active on all further iterations until *stop-test* is true.

  So, to print out the first ten lines of a file, we use the clause:

  ```
  (:range: 1 10 (display record))
  ```

  The colons control whether or not the start and stop lines are processed by the clause. For example:

```
(range   1 5  ...)     Lines   2 3 4
(:range  1 5  ...)     Lines 1 2 3 4
(range:  1 5  ...)     Lines   2 3 4 5
(:range: 1 5  ...)     Lines 1 2 3 4 5
```

A line can trigger both tests, either simultaneously starting and stopping an active region, or simultaneously stopping one and starting a new one, so ranges can abut seamlessly.

- `(else` $body_1$ $body_2$ `...)`
  If no other clause has executed since the top of the loop, or since the last `else` clause, this clause executes.

- `(`*test* `=>` *exp*`)`
  If evaluating `test` produces a true value, apply *exp* to that value. If *test* is a regular expression, then *exp* is applied to the match data structure returned by the regexp match routine.

- `(after` $body_1$ `...)`
  This clause executes when the loop encounters EOF. The body forms execute in the scope of the state vars and the record-count var, if there are any. The value of the last body form is the value of the entire awk form.

  If there is no `after` clause, `awk` returns the loop's state variables as multiple values.

8.2.1 Examples

Here are some examples of `awk` being used to process various types of input stream.

```
(define $ nth)        ; Saves typing.

;;; Print out the name and home-directory of everyone in /etc/passwd:
(let ((read-passwd (field-reader (infix-splitter ":" 7))))
  (call-with-input-file "/etc/passwd"
    (lambda (port)
      (awk (read-passwd port) (record fields) ()
        (#t (format #t "~a's home directory is ~a~%"
                    ($ fields 0)
                    ($ fields 5)))))))
```

```
;;; Print out the user-name and home-directory of everyone whose
;;; name begins with "S"
(let ((read-passwd (field-reader (infix-splitter ":" 7))))
  (call-with-input-file "/etc/passwd"
    (lambda (port)
      (awk (read-passwd port) (record fields) ()
        ((: bos "S")
         (format #t "~a's home directory is ~a~%"
                    ($ fields 0)
                    ($ fields 5)))))))


;;; Read a series of integers from stdin. This expression evaluates
;;; to the number of positive numbers that were read. Note our
;;; "record-reader" is the standard Scheme READ procedure.
(awk (read) (i)   ((npos 0))
  ((> i 0) (+ npos 1)))


;;; Filter -- pass only lines containing my name.
(awk (read-line) (line) ()
  ("Olin" (display line) (newline)))


;;; Count the number of non-comment lines of code in my Scheme source.
(awk (read-line) (line) ((nlines 0))
  ((: bos (* white) ";")  nlines)         ; A comment line.
  (else                   (+ nlines 1)))  ; Not a comment line.


;;; Read numbers, counting the evens and odds.
(awk (read) (val) ((evens 0) (odds 0))
  ((> val 0) (display "pos ")  (values evens odds)) ; Tell me about
  ((< val 0) (display "neg ")  (values evens odds)) ; sign, too.
  (else      (display "zero ") (values evens odds))

  ((even? val) (values (+ evens 1) odds))
  (else        (values evens       (+ odds 1))))


;;; Determine the max length of all the lines in the file.
(awk (read-line) (line) ((max-len 0))
  (#t (max max-len (string-length line))))
```

```
;;; (This could also be done with PORT-FOLD:)
(port-fold (current-input-port) read-line
           (lambda (line maxlen) (max (string-length line) maxlen))
           0)


;;; Print every line longer than 80 chars.
;;; Prefix each line with its line #.
(awk (read-line) (line) lineno ()
  ((> (string-length line) 80)
   (format #t "~d: ~s~%" lineno line)))


;;; Strip blank lines from input.
(awk (read-line) (line) ()
  ((~ white)   (display line) (newline)))


;;; Sort the entries in /etc/passwd by login name.
(for-each (lambda (entry) (display (cdr entry)) (newline))          ; Out
          (sort (lambda (x y) (string<? (car x) (car y)))           ; Sort
                (let ((read (field-reader (infix-splitter ":" 7)))) ; In
                  (awk (read) (line fields) ((ans '()))
                    (#t (cons (cons ($ fields 0) line) ans))))))


;;; Prefix line numbers to the input stream.
(awk (read-line) (line) lineno ()
  (#t (format #t "~d:\t~a~%" lineno line)))
```

8.3 Backwards compatibility

Previous scsh releases provided an awk form with a different syntax, designed around regular expressions written in Posix notation as strings, rather than SREs.

This form is still available in a separate module for old code. It'll be documented in the next release of this manual. Dig around in the sources for it.

Chapter 9

Miscellaneous routines

9.1 Integer bitwise ops

`(arithmetic-shift` *i j*`)` $\longrightarrow$ *integer* procedure
`(bitwise-and` *i j*`)` $\longrightarrow$ *integer* procedure
`(bitwise-ior` *i j*`)` $\longrightarrow$ *integer* procedure
`(bitwise-not` *i*`)` $\longrightarrow$ *integer* procedure
`(bitwise-xor` *i j*`)` $\longrightarrow$ *integer* procedure

These operations operate on integers representing semi-infinite bit strings, using a 2's-complement encoding.

`arithmetic-shift` shifts *i* by *j* bits. A left shift is $j > 0$; a right shift is $j < 0$.

9.2 List procedures

`(nth` *list i*`)` $\longrightarrow$ *object* procedure

Returns the i^{th} element of *list*. The first element (the car) is `(nth` *list* `0)`, the second element is `(nth` *list* `1)`, and so on.

This procedure is provided as it is useful for accessing elements from the lists returned by the field-readers (chapter 8).

9.3 Top level

`(repl)` $\longrightarrow$ *undefined* procedure

This runs a Scheme 48 read-eval-print loop, reading forms from the current input port, and writing their values to the current output port.

If you wish to try something dangerous, and want to be able to recover your shell state, you can fork off a subshell with the following form:

```
(run (begin (repl)))
```

. . . or, rephrased for the proceduralists:

```
(wait (fork repl))
```

Chapter 10

Running scsh

Scsh is currently implemented on top of Scheme 48, a freely-available Scheme implementation written by Jonathan Rees and Richard Kelsey. Scheme 48 uses a byte-code interpreter for good code density, portability and medium efficiency. It is R4RS. It also has a module system designed by Jonathan Rees.

Scsh's design is not Scheme 48 specific, although the current implementation is necessarily so. Scsh is intended to be implementable in other Scheme implementations. The Scheme 48 virtual machine that scsh uses is a specially modified version; standard Scheme 48 virtual machines cannot be used with the scsh heap image.

There are several different ways to invoke scsh. You can run it as an interactive Scheme system, with a standard read-eval-print interaction loop. Scsh can also be invoked as the interpreter for a shell script by putting a "`#!/usr/local/bin/scsh -s`" line at the top of the shell script.

Descending a level, it is also possible to invoke the underlying virtual machine byte-code interpreter directly on dumped heap images. Scsh programs can be pre-compiled to byte-codes and dumped as raw, binary heap images. Writing heap images strips out unused portions of the scsh runtime (such as the compiler, the debugger, and other complex subsystems), reducing memory demands and saving loading and compilation times. The heap image format allows for an initial `#!/usr/local/lib/scsh/scshvm` trigger on the first line of the image, making heap images directly executable as another kind of shell script.

Finally, scsh's static linker system allows dumped heap images to be compiled to a raw Unix a.out(5) format, which can be linked into the text section of the vm binary. This produces a true Unix executable binary file. Since the byte codes comprising the program are in the file's text section, they are not traced or copied by the garbage collector, do not occupy space in the vm's heap, and do not need to be loaded and linked at startup time. This reduces the program's startup time, memory requirements, and paging overhead.

This chapter will cover these various ways of invoking scsh programs.

10.1 Scsh command-line switches

When the scsh top-level starts up, it scans the command line for switches that control its behaviour. These arguments are removed from the command line; the remaining arguments can be accessed as the value of the scsh variable `command-line-arguments`.

10.1.1 Scripts and programs

The scsh command-line switches provide sophisticated support for the authors of shell scripts and programs; they also allow the programmer to write programs that use the Scheme 48 module system.

There is a difference between a *script*, which performs its action *as it is loaded*, and a *program*, which is loaded/linked, and then performs its action by having control transferred to an entry point (*e.g.*, the `main()` function in C programs) that was defined by the load/link operation.

A *script*, by the above definition, cannot be compiled by the simple mechanism of loading it into a scsh process and dumping out a heap image—it executes as it loads. It does not have a top-level `main()`-type entry point.

It is more flexible and useful to implement a system as a program than as a script. Programs can be compiled straightforwardly; they can also export procedural interfaces for use by other Scheme packages. However, scsh supports both the script and the program style of programming.

10.1.2 Inserting interpreter triggers into scsh programs

When Unix tries to execute an executable file whose first 16 bits are the character pair "`#!`", it treats the file not as machine-code to be directly executed by the native processor, but as source code to be executed by some interpreter. The interpreter to use is specified immediately after the "`#!`" sequence on the first line of the source file (along with one optional initial argument). The kernel reads in the name of the interpreter, and executes that instead. The interpreter is passed the source filename as its first argument, with the original arguments following. Consult the Unix man page for the `exec` system call for more information.

Scsh allows Scheme programs to have these triggers placed on their first line. Scsh treats the character sequence "`#!`" as a block-comment sequence,[1] and skips all following characters until it reads the comment-terminating sequence newline/exclamation-point/sharp-sign/newline (*i.e.*, the sequence "`!#`" occurring on its own line).

In this way, the programmer can arrange for an initial

```
#!/usr/local/bin/scsh -s
!#
```

header appearing in a Scheme program to be ignored when the program is loaded into scsh.

10.1.3 Module system

Scsh uses the Scheme 48 module system, which defines *packages*, *structures*, and *interfaces*.

Package A package is an environment—that is, a set of variable/value bindings. You can evaluate Scheme forms inside a package, or load a file into a package. Packages export sets of bindings; these sets are called *structures*.

Structure A structure is a named view on a package—a set of bindings. Other packages can *open* the structure, importing its bindings into their environment. Packages can provide more than one structure, revealing different portions of the package's environment.

Interface An interface is the "type" of a structure. An interface is the set of names exported by a structure. These names can also be marked with other static information (*e.g.*, advisory type declarations, or syntax information).

More information on the the Scheme 48 module system can be found in the file `module.ps` in the `doc` directory of the Scheme 48 and scsh releases.

Programming Scheme with a module system is different from programming in older Scheme implementations, and the associated development problems are consequently different. In Schemes that lack modular abstraction mechanisms, everything is accessible; the major problem is preventing name-space conflicts. In Scheme 48, name-space conflicts vanish; the major problem is that not all bindings are accessible from every place. It takes a little extra work to specify what packages export which values.

[1]Why a block-comment instead of an end-of-line delimited comment? See the section on meta-args.

It may take you a little while to get used to the new style of program development. Although scsh can be used without referring to the module system at all, we recommend taking the time to learn and use it. The effort will pay off in the construction of modular, factorable programs.

Module warning

Programmers who open both the `scheme` and `scsh` structures in their own packages should make sure to always put the `scsh` reference first.

Do this:

```
(define-structure web-server
  (open scsh
        scheme
        net-hax
        ⋮)
  (file web))
```

Open `scsh` before `scheme`.

Not this:

```
(define-structure web-server
  (open scheme
        scsh
        net-hax
        ⋮)
  (file web))
```

Not `scsh` after `scheme`.

Ordering the two packages like this is necessary because scsh overrides some of the standard R4RS Scheme definitions exported by the `scheme` package with its own definitions. For example, scsh's versions of the R4RS I/O functions such as `display` and `write` take integer file descriptors as arguments, as well as Scheme ports. If you open the `scheme` structure before the `scsh` structure, you'll get the standard Scheme 48 definitions, which is not what you want.

10.1.4 Switches

The scsh top-level takes command-line switches in the following format:

`scsh` [*meta-arg*] [$switch_i$ ...] [*end-option* arg_1 ... arg_n]

where

meta-arg:	\ *script-file-name*	
switch:	`-e` *entry-point*	Specify top-level entry-point.
	`-o` *structure*	Open structure in current package.
	`-m` *structure*	Switch to package.
	`-n` *new-package*	Switch to new package.
	`-lm` *module-file-name*	Load module into config package.
	`-l` *file-name*	Load file into current package.
	`-dm`	Do script module.
	`-ds`	Do script.
end-option:	`-s` *script*	
	`-sfd` *num*	
	`-c` *exp*	
	`--`	

These command-line switches essentially provide a little linker language for linking a shell script or a program together with Scheme 48 modules. The command-line processor serially opens structures and loads code into a given package. Switches that side-effect a package operate on a particular "current" package; there are switches to change this package. (These switches provide functionality equivalent to the interactive `,open` `,load` `,in` and `,new` commands.) Except where indicated, switches specify actions that are executed in a left-to-right order. The initial current package is the user package, which is completely empty and opens (imports the bindings of) the R4RS and scsh structures.

If the Scheme process is started up in an interactive mode, then the current package in force at the end of switch scanning is the one inside which the interactive read-eval-print loop is started.

The command-line switch processor works in two passes: it first parses the switches, building a list of actions to perform, then the actions are performed serially. The switch list is terminated by one of the *end-option* switches. The arg_i arguments occurring after an end-option switch are passed to the scsh program as the value of `command-line-arguments` and the tail of the list returned by `(command-line)`. That is, an *end-option* switch separates switches that control the scsh "machine" from the actual arguments being passed to the scsh program that runs on that machine.

The following switches and end options are defined:

- `-o` *struct*
  Open the structure in the current package.

- `-n` *package*
  Make and enter a new package. The package has an associated structure named *package* with an empty export list. If *package* is the string "`#f`", the new package is anonmyous, with no associated named structure.

  The new package initially opens no other structures, not even the R4RS bindings. You must follow a "`-n foo`" switch with "`-o scheme`" to access the standard identifiers such as `car` and `define`.

- `-m` *struct*
  Change the current package to the package underlying structure *struct*. (The `-m` stands for "module.")

- `-lm` *module-file-name*
  Load the specified file into scsh's config package — the file must contain source written in the Scheme 48 module language ("load module"). Does not alter the current package.

- `-l` *file-name*
  Load the specified file into the current package.

- `-c` *exp*
  Evaluate expression *exp* in the current package and exit. This is called `-c` after a common shell convention (see sh and csh). The expression is evaluated in the the current package (and hence is affected by `-m`'s and `-n`'s.)

  When the scsh top-level constructs the scsh command-line in this case, it takes `"scsh"` to be the program name. This switch terminates argument scanning; following args become the tail of the command-line list.

- `-e` *entry-point*
  Specify an entry point for a program. The *entry-point* is a variable that is taken from the current package in force at the end of switch evaluation. The entry point does not have to be exported by the package in a structure; it can be internal to the package. The top level passes control to the entry point by applying it to the command-line list (so programs executing in private packages can reference their command-line arguments without opening the `scsh` package to access the `(command-line)` procedure). Note that, like the list returned by the `(command-line)` procedure, the list passed to the entry point includes the name of the program being executed (as the first element of the list), not just the arguments to the program.

  A `-e` switch can occur anywhere in the switch list, but it is the *last* action performed by switch scanning if it occurs. (We violate ordering here as the shell-script `#!` mechanism prevents you from putting the *-e* switch last, where it belongs.)

- `-s` *script*
  Specify a file to load. A `-ds` (do-script) or `-dm` (do-module) switch occurring earlier in the switch list gives the place where the script should be loaded. If there is no `-ds` or `-dm` switch, then the script is loaded at the end of switch scanning, into the module that is current at the end of switch scanning.

  We use the `-ds` switch to violate left-to-right switch execution order as the `-s` switch is *required* to be last (because of the `#!` machinery), independent of when/where in the switch-processing order it should be loaded.

  When the scsh top-level constructs the scsh command-line in this case, it takes *script* to be the program name. This switch terminates switch parsing; following args are ignored by the switch-scanner and are passed through to the program as the tail of the command-line list.

- `-sfd` *num*
  Loads the script from file descriptor *num*. This switch is like the `-s` switch, except that the script is loaded from one of the process' open input file descriptors. For example, to have the script loaded from standard input, specify `-sfd 0`.

- `--`
  Terminate argument scanning and start up scsh in interactive mode. If the argument list just runs out, without either a terminating `-s` or `--` arg, then scsh also starts up in interactive mode, with an empty `command-line-arguments` list (for example, simply entering `scsh` at a shell prompt with no args at all).

  When the scsh top-level constructs the scsh command-line in this case, it takes `"scsh"` to be the program name. This switch terminates switch parsing; following args are ignored by the switch-scanner and are passed through to the program as the tail of the command-line list.

- `-ds`
  Specify when to load the script ("do-script"). If this switch occurs, the switch list *must* be terminated by a `-s` *script* switch. The script is loaded into the package that is current at the `-ds` switch.

- `-dm`
  As above, but the current module is ignored. The script is loaded into the `config` package ("do-module"), and hence must be written in the Scheme 48 module language. This switch doesn't affect the current module—after executing this switch, the current module is the same as as it was before.

  This switch is provided to make it easy to write shell scripts in the Scheme 48 module language.

10.1.5 The meta argument

The scsh switch parser takes a special command-line switch, a single backslash called the "meta-argument," which is useful for shell scripts. If the initial command-line argument is a "\" argument, followed by a filename argument *fname*, scsh will open the file *fname* and read more arguments from the second line of this file. This list of arguments will then replace the "\" argument—*i.e.*, the new arguments are inserted in front of *fname*, and the argument parser resumes argument scanning. This is used to overcome a limitation of the `#!` feature: the `#!` line can only specify a single argument after the interpreter. For example, we might hope the following scsh script, `ekko`, would implement a simple-minded version of the Unix `echo` program:

```
#!/usr/local/bin/scsh -e main -s
!#
(define (main args)
  (map (λ (arg) (display arg) (display " "))
       (cdr args))
  (newline))
```

The idea would be that the command

```
ekko Hi there.
```

would by expanded by the `exec(2)` kernel call into

```
/usr/local/bin/scsh -e main -s ekko Hi there.
```

In theory, this would cause scsh to start up, load in file `ekko`, call the entry point on the command-line list

```
(main '("ekko" "Hi" "there."))
```

and exit.

Unfortunately, the Unix `exec(2)` syscall's support for scripts is not very general or well-designed. It will not handle multiple arguments; the `#!` line is usually required to contain no more than 32 characters; it is not recursive. If these restrictions are violated, most Unix systems will not provide accurate error reporting, but either fail silently, or simply incorrectly implement the desired functionality. These are the facts of Unix life.

In the `ekko` example above, our `#!` trigger line has three arguments ("`-e`", "`main`", and "`-s`"), so it will not work. The meta-argument is how we work around this problem. We must instead invoke the scsh interpreter with the single \ argument, and put the rest of the arguments on line two of the program. Here's the correct program:

```
#!/usr/local/bin/scsh \
-e main -s
!#
(define (main args)
  (map (λ (arg) (display arg) (display " "))
       (cdr args))
  (newline))
```

Now, the invocation starts as

```
ekko Hi there.
```

and is expanded by exec(2) into

```
/usr/local/bin/scsh \ ekko Hi there.
```

When scsh starts up, it expands the "\" argument into the arguments read from line two of `ekko`, producing this argument list:

```
-e main -s ekko Hi there.
       ↑
Expanded from \ ekko
```

With this argument list, processing proceeds as we intended.

Secondary argument syntax

Scsh uses a very simple grammar to encode the extra arguments on the second line of the scsh script. The only special characters are space, tab, newline, and backslash.

- Each space character terminates an argument. This means that two spaces in a row introduce an empty-string argument.
- The tab character is not permitted (unless you quote it with the backslash character described below). This is to prevent the insidious bug where you believe you have six space characters, but you really have a tab character, and *vice-versa*.
- The newline character terminates an argument, like the space character, and also terminates the argument sequence. This means that an empty line parses to the singleton list whose one element is the empty string: `("")`. The grammar doesn't admit the empty list.

- The backslash character is the escape character. It escapes backslash, space, tab, and newline, turning off their special functions, and allowing them to be included in arguments. The ANSI C escape sequences (`\b`, `\n`, `\r` and `\t`) are also supported; these also produce argument-constituents—`\n` doesn't act like a terminating newline. The escape sequence *nnn* for *exactly* three octal digits reads as the character whose ASCII code is *nnn*. It is an error if backslash is followed by just one or two octal digits: `\3Q` is an error. Octal escapes are always constituent chars. Backslash followed by other chars is not allowed (so we can extend the escape-code space later if we like).

You have to construct these line-two argument lines carefully. In particular, beware of trailing spaces at the end of the line—they'll give you extra trailing empty-string arguments. Here's an example:

```
#!/bin/interpreter \
foo bar  quux\ yow
```

would produce the arguments

```
("foo" "bar" "" "quux yow")
```

10.1.6 Examples

- `scsh -dm -m myprog -e top -s myprog.scm`
  Load `myprog.scm` into the `config` package, then shift to the `myprog` package and call `(top '("myprog.scm"))`, then exit. This sort of invocation is typically used in `#!` script lines (see below).

- `scsh -c '(display "Hello, world.")'`
  A simple program.

- `scsh -o bigscheme`
  Start up interactively in the user package after opening structure `bigscheme`.

- `scsh -o bigscheme -- Three args passed`
  Start up interactively in the user package after opening `bigscheme`. The `command-line-args` variable in the scsh package is bound to the list `("Three" "args" "passed")`, and the `(command-line)` procedure returns the list `("scsh" "Three" "args" "passed")`.

- Program `ekko`
  This shell script, called `ekko`, implements a version of the Unix `echo` program:

```
#!/usr/local/bin/scsh -s
!#
(for-each (λ (arg) (display arg) (display " "))
          command-line-args)
```

Note this short program is an example of a *script*—it executes as it loads. The Unix rule for executing `#!` shell scripts causes

```
ekko Hello, world.
```

to expand as

```
/usr/local/bin/scsh -s ekko Hello, world.
```

- Program `ekko`
  This is the same program, *not* as a script. Writing it this way makes it possible to compile the program (and then, for instance, dump it out as a heap image).

  ```
  #!/usr/local/bin/scsh \
  -e top -s
  !#
  (define (top args)
    (for-each (λ (arg) (display arg) (display " "))
              (cdr args)))
  ```

  The `exec(2)` expansion of the `#!` line together with the scsh expansion of the "`\ ekko`" meta-argument (see section 10.1.5) gives the following command-line expansion:

  ```
  ekko Hello, world.
      ⟹  /usr/local/bin/scsh \ ekko         Hello, world.
      ⟹  /usr/local/bin/scsh -e top -s ekko Hello, world.
  ```

- Program `sort`
  This is a program to replace the Unix `sort` utility—sorting lines read from stdin, and printing the results on stdout. Note that the source code defines a general sorting package, which is useful (1) as a Scheme module exporting sort procedures to other Scheme code, and (2) as a standalone program invoked from the `top` procedure.

```
#!/usr/local/bin/scsh \
-dm -m sort-toplevel -e top -s
!#

;;; This is a sorting module. TOP procedure exports
;;; the functionality as a Unix program akin to sort(1).
(define-structures ((sort-struct (export sort-list
                                         sort-vector!))
                    (sort-toplevel (export top)))
  (open scheme)

  (begin (define (sort-list elts <=) ...)
         (define (sort-vec! vec <=)  ...)

         ;; Parse the command line and
         ;; sort stdin to stdout.
         (define (top args)
            ...)))
```

The expansion below shows how the command-line scanner (1) loads the config file `sort` (written in the Scheme 48 module language), (2) switches to the package underlying the `sort-toplevel` structure, (3) calls `(top '("sort" "foo" "bar"))` in the package, and finally (4) exits.

```
sort foo bar
⟹  /usr/local/bin/scsh \ sort                           foo bar
⟹  /usr/local/bin/scsh -dm -m sort-toplevel -e top -s sort foo bar
```

An alternate method would have used a

```
-n #f -o sort-toplevel
```

sequence of switches to specify a top-level package.

Note that the sort example can be compiled into a Unix program by loading the file into an scsh process, and dumping a heap with top-level `top`. Even if we don't want to export the sort's functionality as a subroutine library, it is still useful to write the sort program with the module language. The command line design allows us to run this program as either an interpreted script (given the `#!` args in the header) or as a compiled heap image.

10.1.7 Process exit values

Scsh ignores the value produced by its top-level computation when determining its exit status code. If the top-level computation completed with no errors, scsh dies with exit code 0. For example, a scsh process whose top-level is specified by a `-c` *exp* or a `-e` *entry* entry point ignores the value produced by evaluating *exp* and calling *entry*, respectively. If these computations terminate with no errors, the scsh process exits with an exit code of 0.

To return a specific exit status, use the `exit` procedure explicitly, *e.g.*,

```
scsh -c \
  "(exit (status:exit-val (run (| (fmt) (mail shivers)))))"
```

10.2 The scsh virtual machine

To run the Scheme 48 implementation of scsh, you run a specially modified copy of the Scheme 48 virtual machine with a scsh heap image. The scsh binary is actually nothing but a small cover program that invokes the byte-code interpreter on the scsh heap image for you. This allows you to simply start up an interactive scsh from a command line, as well as write shell scripts that begin with the simple trigger

```
#!/usr/local/bin/scsh -s
```

You can also directly execute the virtual machine, which takes its own set of command-line switches.. For example, this command starts the vm up with a 1Mword heap (split into two semispaces):

```
scshvm -o scshvm -h 1000000 -i scsh.image arg1 arg2 ...
```

The vm peels off initial vm arguments up to the `-i` heap image argument, which terminates vm argument parsing. The rest of the arguments are passed off to the scsh top-level. Scsh's top-level removes scsh switches, as discussed in the previous section; the rest show up as the value of `command-line-arguments`.

Directly executing the vm can be useful to specify non-standard switches, or invoke the virtual machine on special heap images, which can contain precompiled scsh programs with their own top-level procedures.

10.2.1 VM arguments

The vm takes arguments in the following form:

`scshvm [`*meta-arg*`] [`*vm-options*$^+$`] [`*end-option scheme-args*`]`

where

meta-arg:	`\` *filename*
vm-option:	`-h` *heap-size-in-words* `-s` *stack-size-in-words* `-o` *object-file-name*
end-option:	`-i` *image-file-name* `--`

The vm's meta-switch "`\` *filename*" is handled the same as scsh's meta-switch, and serves the same purpose.

VM options

The `-o` *object-file-name* switch tells the vm where to find relocation information for its foreign-function calls. Scsh will use a pre-compiled default if it is not specified. Scsh *must* have this information to run, since scsh's syscall interfaces are done with foreign-function calls.

The `-h` and `-s` options tell the vm how much space to allocate for the heap and stack. The heap size value is the total number of words allocated for the heap; this space is then split into two semi-spaces for Scheme 48's stop-and-copy collector.

End options

End options terminate argument parsing. The `-i` switch is followed by the name of a heap image for the vm to execute. The *image-file-name* string is also taken to be the name of the program being executed by the VM; this name becomes the head of the argument list passed to the heap image's top-level entry point. The tail of the argument list is constructed from all following arguments.

The `--` switch terminates argument parsing without giving a specific heap image; the vm will start up using a default heap (whose location is compiled into the vm). All the following arguments comprise the tail of the list passed off to the heap image's top-level procedure.

Notice that you are not allowed to pass arguments to the heap image's top-level procedure (*e.g.*, scsh) without delimiting them with `-i` or `--` flags.

10.2.2 Inserting interpreter triggers into heap images

Scheme 48's heap image format allows for an informational header: when the vm loads in a heap image, it ignores all data occurring before the first control-L character (ASCII 12). This means that you can insert a "`#!`" trigger line into a heap image, making it a form of executable "shell script." Since the vm requires multiple arguments to be given on the command line, you must use the meta-switch. Here's an example heap-image header:

```
#!/usr/local/lib/scsh/scshvm \
-o /usr/local/lib/scsh/scshvm -i
... Your heap image goes here ...
```

10.2.3 Inserting a double-level trigger into Scheme programs

If you're a nerd, you may enjoy doing a double-level machine shift in the trigger line of your Scheme programs with the following magic:

```
#!/usr/local/lib/scsh/scshvm \
-o /usr/local/lib/scsh/scshvm -i /usr/local/lib/scsh/scsh.image -s
!#
... Your Scheme program goes here ...
```

10.3 Compiling scsh programs

Scsh allows you to create a heap image with your own top-level procedure. Adding the pair of lines

```
#!/usr/local/lib/scsh/scshvm \
-o /usr/local/lib/scsh/scshvm -i
```

to the top of the heap image will turn it into an executable Unix file.

You can create heap images with the following two procedures.

(`dump-scsh-program` *main fname*) $\longrightarrow$ *undefined* procedure

This procedure writes out a scsh heap image. When the heap image is executed by the Scheme 48 vm, it will call the *main* procedure, passing it the vm's argument list. When `main` returns an integer value i, the vm exits with exit status i. The Scheme vm will parse command-line switches as described in section 10.2.1; remaining arguments form the tail of the command-line list that is passed to `main`. (The head of the list is the name of the program being executed by the vm.) Further argument parsing (as described for scsh in section 10.1.4) is not performed.

The heap image created by `dump-scsh-program` has unused code and data pruned out, so small programs compile to much smaller heap images.

(`dump-scsh` *fname*) $\longrightarrow$ *undefined* procedure

This procedure writes out a heap image with the standard scsh top-level. When the image is resumed by the vm, it will parse and execute scsh command-line switches as described in section 10.1.4.

You can use this procedure to write out custom scsh heap images that have specific packages preloaded and start up in specific packages.

Unfortunately, Scheme 48 does not support separate compilation of Scheme files or Scheme modules. The only way to compile is to load source and then dump out a heap image. One occasionally hears rumours that this is being addressed by the Scheme 48 development team.

10.4 Statically linking heap images

The static heap linker converts a Scheme 48 bytecode image contained in a .image file to a C representation. This C code is then compiled and linked in with a virtual machine, producing a single executable. Some of the benefits are:

- Instantaneous start-up time.
- Improved paging; scsh images can be shared between different processes.
- Vastly reduced GC copying—the whole initial image is moved out of the heap, and neither traced nor copied.
- Result program no longer depends on the filesystem for its initial image.

The static heap linker takes arguments in the following form:

`scsh-hlink` *image executable* [*option* ...]

It reads in the heap image *image*, translates it into C code, compiles the C code, and links it against the scsh vm, producing the standalone binary file *executable*.

Each C file represents part of the heap image as a constant C `long` vector that looks something like this:

```
const  long p116[]={0x882,0x24,0x19,
                  0x882,(long)(&p19[785])+7,(long)(&p119[125])+7,
                  0x882,(long)(&p119[128])+7,(long)(&p119[131])+7,
                  0x882,(long)(&p102[348])+7,(long)(&p3[114])+7,
                  0xfc2,0x2030200,0x7100209,0x1091002,0x1c075a,
                  0x882,(long)(&p29[1562])+7,(long)(&p119[137])+7,
                  0x882,(long)(&p78[692])+7,(long)(&p119[140])+7,
                      .
                      .
                      .
                  };
```

Translating to a C declaration gives us freedom from the various object-file formats.[2] Note that the const declaration allows the compiler to put this array in the text pages of the executable. The heap is split into parts because many C compilers cannot handle multi-megabyte initialised vector declarations.

The allowed options to the heap linker are:

- `--temp` *dir*
  The temporary directory to hold .c and .o files. The default is typically configured to be `/usr/tmp`, and can be overridden by the environment variable `TMPDIR`. Example:

  ```
  --temp /tmp
  ```

- `--cc` *command*
  The command to run the C compiler. The default can be overridden by the environment variable `CC`. Example:

  ```
  --cc "gcc -g -O"
  ```

- `--ld` *command*
  The arguments to run the C compiler as a linker. The default can be overridden by the environment variable `LDFLAGS`. Example:

  ```
  --ld "-Wl,-E"
  ```

- `--libs` *libs*
  The libraries needed to link the VM and heap. The default can be overridden by the environment variable `LIBS`. Example:

  ```
  --libs "-ldld -lld -lm"
  ```

Be warned that the current heap linker has many shortcomings.

- It is extremely slow. Really, really slow. Translating the standard scsh heap image into a standalone binary takes well over an hour on a 40M-b/133Mhz Pentium system. A memory-starved 486 could take all night.
- It cannot be applied to itself. The current implementation works by replacing some of the heap-dumping code. This means you cannot load the heap-linker code into a scsh system and subsequently use `dump-scsh-program` to create a heap-linker heap image.
- The interface leaves a lot to be desired.
  - It requires the heap image to be referenced by a file-name; the linker will not allow you to feed it the input heap image on a port.

[2] This idea is due to Jonathan Rees.

- The heap-image is linked against the vm contained in
  `/usr/local/lib/scsh/libscshvm.a`
  This is wired in at the time scsh is installed on your system.
- There is no Scheme procedural interface.

- The program produced uses the default VM argv parser `process_args` from the scsh source file `main.c` to process the command line before handing it off to the heap image's top-level procedure. This is not what you want for many programs.

  The system needs to be changed to allow users to override this default with their own VM argument parsers.

- A possible problem is the Unix limits on the number of command line arguments. The heap-linker calls the C linker with a large number of object files. Its conceivable that on some Unix systems this could fail now or if scsh grows in the future. The solution could be to create library archives of a few dozen files and then link the result few dozen library archives to make the executable.

In spite of these many shortcomings, we are providing the static linker as it stands in this release so that people may get some experience with it.

Here is an example of how one might use the heap linker:

```
scsh-hlink scsh.image fastscsh
```

We'd love it if someone would dive into the source and improve it.

10.5 Standard file locations

Because the scshvm binary is intended to be used for writing shell scripts, it is important that the binary be installed in a standard place, so that shell scripts can dependably refer to it. The standard directory for the scsh tree should be `/usr/local/lib/scsh/`. Whenever possible, the vm should be located in

```
/usr/local/lib/scsh/scshvm
```

and a scsh heap image should be located in

```
/usr/local/lib/scsh/scsh.image
```

The top-level scsh program should be located in

```
/usr/local/lib/scsh/scsh
```

with a symbolic link to it from

```
/usr/local/bin/scsh
```

The Scheme 48 image format allows heap images to have `#!` triggers, so `scsh.image` should have a `#!` trigger of the following form:

```
#!/usr/local/lib/scsh/scshvm \
-o /usr/local/lib/scsh/scshvm -i
... heap image goes here ...
```

Chapter 11

Changes from previous releases

11.1 Changes from the previous release

This section details changes that have been made in scsh since the previous release.

Scsh is now much more robust. All known bugs have been fixed. There have been many improvements and extensions made. These new features and changes are listed below, in no particular order; the relevant sections of the manual give the full details.

Scsh now supports complete POSIX, including signal handlers. Early autoreaping of child processes is now handled by a `SIGCHLD` signal handler, so children are reaped as early as possible with no user intervention required.

A functional static heap linker is included in this release. It is ugly, limited in functionality, and extremely slow, but it works. It can be used to build scsh binaries that start up instantly.

The regular expression system has been sped up. Regular-expression compilation is now provided, and the `awk` macro has been rewritten to pre-compile regexps used in rules outside the loop. It is still, however, slower than it should be.

Execing programs should be faster in this release, since we now use the `CLOEXEC` status bit to get automatic closing of unrevealed port file descriptors.

scm's floating point support was inadvertently omitted from the last release. It has been reinstated.

There is now a new command-line switch, `-sfd` *num*, which causes scsh to read its script from file descriptor *num*.

11.2 Changes from the penultimate release

This section details changes that have been made in scsh since the penultimate release.

Scsh is now much more robust. All known bugs have been fixed. There have been many improvements and extensions made. We have also made made some incompatible changes.

The sections below briefly describe these new features and changes; the relevant sections of the manual give the full details.

11.2.1 New features

This release incorporates several new features into scsh.

- Control of buffered I/O
  Scsh now allows you to control the buffering policy used for doing I/O on a Scheme port.

- Here-strings
  Scsh now has a new lexical feature, `#<<`, that provides the ability to enter long, multi-line string constants in scsh programs. Such a string is called a "here string," by analogy to the common shell "here document" << redirection.

- Delimited readers and read-line
  Scsh now has a powerful set of delimited readers. These can be used to read input delimited by a newline character (`read-line`), a blank line (`read-paragraph`), or the occurrence of any character in an arbitrary set (`read-delimited`).

  While these procedures can be applied to any Scheme input port, there is native-code support for performing delimited reads on Unix input sources, so doing block input with these procedures should be much faster than the equivalent character-at-a-time Scheme code.

- New system calls
  With the sole exception of signal handlers, scsh now has all of POSIX. This release introduces

  - `select`,
  - full terminal device control,

- support for pseudo-terminal "pty" devices,
- file locking,
- process timing,
- `set-file-times`,
- `seek` and `tell`.

Note that having `select`, pseudo-terminals, and tty device control means that it is now possible to implement interesting network protocols, such as telnet servers and clients, directly in Scheme.

- New command-line switches
There is a new set of command-line switches that make it possible to write shell scripts using the Scheme 48 module system. Scripts can use the new command-line switches to open dependent modules and load dependent source code. Scripts can also be written in the Scheme 48 module language, which allows you to use it both as a standalone shell script, and as a code module that can be loaded and used by other Scheme programs.

- Static heap linking
There is a new facility that allows you to compile a heap image to a `.o` file that can be linked with the scsh virtual machine. This produces a standalone executable binary, makes startup time near-instantaneous, and greatly improves memory performance—the initial heap image is placed in the process' text pages, where it is shared by different scsh processes, and does not occupy space in the run-time heap.

    Oops: The static heap linker was not documented and installed in time for this release.

11.2.2 Incompatible improvements

Some features of scsh have been improved in ways that are not backwards-compatible with previous releases. These changes should not affect most code; however, please note the changes and modify your code accordingly.

- New process-object data-type returned by `fork`
Previous releases were prone to fill up the kernel's process table if a program forked large numbers of processes and subsequently failed to use `wait` to reclaim the entries in the kernel's process table. (This is a problem in standard C environments, as well.)

  Scsh 0.4 introduces a new mechanism for automatically managing subprocesses. Processes are no longer represented by an integer process id, which is impossible to garbage-collect, but by an abstract process data type that encapsulates the process id. All processes are represented using the new data structures; see the relevant section of the manual for further details.

- Better stdio/current-port synchronisation
  The `(begin ...)` process form now does a `stdio->stdports` call before executing its body. This means that the Scheme code in the body "sees" any external redirections. For example, it means that if a `begin` form in the middle of a pipeline performs I/O on the current input and output ports, it will be communicating with its upstream and downstream pipes. *E.g.*, this code works as intended without the need for explicit synchronisation:

```
(run (| (gunzip)
        ;; Kill line 1 and insert doubled-sided
        ;; code at head of Postscript.
        (begin (read-line) ; Eat first line.
               (display "%!PS-Adobe-2.0\\n")
               (display "statusdict /setduplexmode known ")
               (display "{statusdict begin true ")
               (display "setduplexmode end} if\n")
               (exec-epf (cat)))
        (lpr))
     (< paper.ps))
```

  Arranging for the `begin` process form to synchronise the current I/O ports with stdio means that all process forms now see their epf's redirections.

- `file-match` more robust
  The `file-match` procedure now catches any error condition signalled by a match procedure, and treats it as if the procedure had simply returned `#f`, *i.e.*, match failure. This means `file-match` no longer gets blown out of the water by trying to apply a function like `file-directory?` to a dangling symlink, and other related OS errors.

- Standard input now unbuffered
  Scsh's startup code now makes the initial current input port (corresponding to file descriptor 0) unbuffered. This keeps the shell from "stealing" input meant for subprocesses. However, it does slow down character-at-a-time input processing. If you are writing a program that is tolerant of buffered input, and wish the efficiency gains, you can reset the buffering policy yourself.

- “writeable” now spelled “writable”
  We inconsistently spelled `file-writable?` and `file-not-writable?` in the manual and the implementation. We have now standardised on the common spelling “writable” in both. The older bindings still exist in release 0.4, but will go away in future releases.

- `char-set-member?` replaced
  We have de-released the `char-set-member?` procedure. The scsh 0.3 version of this procedure took arguments in the following order:

  (`char-set-member?` *char char-set*)

  This argument order is in accordance with standard mathematical useage (*i.e.*, $x \in S$), and also consistent with the R4RS `member`, `memq` and `memv` procedures. It is, however, exactly opposite from the argument order used by the `char-set-member?` in MIT Scheme's character-set library. If we left things as they were, we risked problems with code ported over from MIT Scheme. On the other hand, changing to conformance with MIT Scheme meant inconsistency with common mathematical notation and other long-standing Scheme procedures. Either way was bound to introduce confusion.

  We've taken the approach of simply removing the `char-set-member?` procedure altogether, and replacing it with a new procedure:

  (`char-set-contains?` *cset char*)

  Note that the argument order is consistent with the name.

- `file-attributes` now `file-info`
  In keeping with the general convention in scsh of naming procedures that retrieve information about system resources `...-info` (*e.g.*, `tty-info`, `user-info`, `group-info`), the `file-attributes` procedure is now named `file-info`.

  We continue to export a `file-attributes` binding for the current release, but it will go away in future releases.

- Renaming of I/O synchronisation procedures
  The (`stdio->stdports` *thunk*) procedure has been renamed `with-stdio-ports*`; there is now a corresponding `with-stdio-ports` special form. The `stdio->stdports` procedure is now a nullary procedure that side-effects the current set of current I/O port bindings.

- New meta-arg line-two syntax
  Scsh now uses a simplified grammar for describing command-line arguments read by the “meta-arg” switch from line two of a shell script. If you were using this feature in previous releases, the three incompatible

changes of which to be aware are: (1) tab is no longer allowed as an argument delimiter, (2) a run of space characters is not equivalent to a single space, (3) empty arguments are written a different way.

11.2.3 Backwards-compatible improvements

Some existing features in scsh have been improved in ways that will not effect existing code.

- Improved error reporting
  Exception handlers that print out error messages and warnings now print their messages on the error output port, instead of the current output port. Previous releases used the current output port, a problem inherited from Scheme 48.

  Previous scsh releases flushed the Scheme 48 debugging tables when creating the standard scsh heap image. This trimmed the size of the heap image, but made error messages much less comprehensible. We now retain the debugging tables. This bloats the heap image up by about 600kb. And worth it, too.

  (We also have some new techniques for eliminating the run-time memory penalty imposed by these large heap images. Scsh's new static-heap technology allows for this data to be linked into the text pages of the vm's binary, where it will not be touched by the GC or otherwise affect the memory system until it is referenced.)

  Finally, scsh now generates more informative error messages for syscall errors. For example, a file-open error previously told you what the error was (*e.g.*, "Permission denied," or "No such file or directory"), but not which file you had tried to open. We've improved this.

- Closing a port twice allowed
  Scsh used to generate an error if you attempted to close a port that had already been closed. This is now allowed. The close procedure returns a boolean to indicate whether the port had already been closed or not.

- Better time precision
  The `time+ticks` procedure now returns sub-second precision on OS's that support it.

- Nicer print-methods for basic data-types
  Scsh's standard record types now print more informatively. For example, a process object includes the process id in its printed representation: the process object for process id 2653 prints as `#{proc 2653}`.

Chapter 12

Todo

There are always many, many improvements and extensions that could be made to scsh. We invite interested hackers to do any of them, and send us the code; we'll put you on the team. Visit the Scheme Underground Web page for more information on good hacks at

`http://www.ai.mit.edu/projects/su/`

Scsh is a tool that lets you write fun programs that do real things in an elegant language; go wild.

- Threads.
- An X gui interface. (Needs threads.)
- A better C function/data-structure interface. This is not easy.
- More network protocols. Telnet and ftp would be the most important.
- An ILU interface.
- An RPC system, with "tail-recursion."
- Interfaces to relational db's. This would be quite useful for Web servers. An s-expression embedding of SQL would be a key design component of such a system, along the lines of scsh's process notation or `awk` notation.
- Port Edwin, and emacs text editor written in MIT Scheme, to scsh. Combine it with scsh's OS interfaces to make a visual shell.
- An `expect` knock-off.
- A `make` replacement, using scsh's process notation in the build rules.
- Manual hacking.

  - The LaTeX hackery needs yet another serious pass. Most importantly, long procedure "declarations" need to be broken across two lines.
  - Soup up the markup processor, and redo manual in markup. Generate LaTeX, HTML, and info versions. Alternatively, persuade some kind soul to hand-port manual to HTML or info.

- Job control, after `jcontrol.scm`
- Better static heap linker.
- Gnu readline lib.

Index

%read-delimited!, 141
temp-file-template, 50
->uid, 62
->username, 62
%exec, 52
%exit, 53
%fork, 53
%fork/pipe, 54
%fork/pipe+, 55
&, 13
&, 13
&&, 20
interrupt/alarm, 65
interrupt/alrm, 65
interrupt/chld, 65
interrupt/cont, 65
interrupt/hup, 65
interrupt/info, 65
interrupt/int, 65
interrupt/io, 65
interrupt/keyboard, 65
interrupt/memory-shortage, 65
interrupt/poll, 65
interrupt/prof, 65
interrupt/pwr, 65
interrupt/quit, 65
interrupt/term, 65
interrupt/tstp, 65
interrupt/urg, 65
interrupt/usr1, 65
interrupt/usr2, 65
interrupt/vtalrm, 65
interrupt/winch, 65
interrupt/xcpu, 65
interrupt/xfsz, 65
signal/abrt, 66
signal/alrm, 65
signal/bus, 66
signal/chld, 65
signal/cont, 65
signal/emt, 66
signal/fpe, 66
signal/hup, 65
signal/ill, 66
signal/info, 65
signal/int, 65
signal/iot, 66
signal/io, 65
signal/kill, 66
signal/pipe, 66
signal/poll, 65
signal/prof, 65
signal/pwr, 65
signal/quit, 65
signal/segv, 66
signal/stop, 66
signal/sys, 66
signal/term, 65
signal/trap, 66
signal/tstp, 65
signal/ttin, 66
signal/ttou, 66
signal/urg, 65
signal/usr1, 65
signal/usr2, 65
signal/vtalrm, 65
signal/winch, 65
signal/xcpu, 65

signal/xfsz, 65

reduce-char-set , 107
reduce-port , 16

absolute-file-name, 105
accept-connection, 94
add-after, 76
add-before, 76
alist->env, 75
alist-compress, 75
alist-delete, 75
alist-update, 75
arg, 62
arg*, 62
argv, 62
arithmetic-shift, 156
ascii->char, 106
ascii-range->char-set, 107
autoreap-policy, 56
awk, 151

become-session-leader, 83
bind-listen-accept-loop, 91
bind-socket, 94
bitwise-and, 156
bitwise-ior, 156
bitwise-not, 156
bitwise-xor, 156

call-terminally, 53
call-with-string-output-port, 26
call/fdes, 30
char->ascii, 106
char-alphabetic?, 110
char-alphanumeric?, 110
char-ascii?, 110
char-blank?, 110
char-control?, 110
char-filter, 20
char-graphic?, 110
char-hex-digit?, 110
char-lower-case?, 110
char-numeric? , 110
char-printing?, 110
char-punctuation?, 110
char-set, 107
char-set-adjoin, 109
char-set-any?, 108
char-set-contains?, 108
char-set-copy, 111
char-set-delete, 109
char-set-difference, 108
char-set-every?, 108
char-set-fold, 107
char-set-for-each, 107
char-set-intersection, 108
char-set-invert, 108
char-set-members, 108
char-set-size, 108
char-set-union, 108
char-set:alphabetic, 109
char-set:alphanumeric, 109
char-set:ascii, 109
char-set:blank, 109
char-set:control, 109
char-set:empty, 109
char-set:full, 109
char-set:graphic, 109
char-set:hex-digit, 109
char-set:lower-case, 109
char-set:numeric, 109
char-set:printing, 109
char-set:punctuation, 109
char-set:upper-case, 109
char-set:whitespace, 109
char-set<=, 107
char-set=, 107
char-set?, 106
char-upper-case?, 110
char-whitespace?, 110
chars->char-set, 107
chdir, 60
clean-up-cres, 138
close, 25
close-after, 24
close-socket, 91

command-line, 62
command-line-arguments, 62
connect-socket, 94
control-tty-file-name, 83
copy-tty-info, 80
cpu-ticks/sec, 61
create-directory, 40
create-fifo, 40
create-hard-link, 40
create-socket, 91
create-socket-pair, 91
create-symlink, 40
create-temp-file, 49
cwd, 60

date, 69
date, 71
date->string, 72
define-record, 7
delete-directory, 40
delete-file, 40
delete-filesys-object, 40
directory-as-file-name, 101
directory-files, 45
disable-tty-char, 79
drain-tty, 81
dump-scsh, 172
dump-scsh-program, 172
dup, 31
dup->fdes, 31
dup->inport, 31
dup->outport, 31

enabled-interrupts, 67
env->alist, 74
errno-error, 21
error-output-port, 24
exec, 52
exec-epf, 13
exec-epf, 13
exec-path, 52
exec-path-list, 77
exec-path-search, 52
exec-path/env, 52
exec/env, 52
exit, 53
expand-file-name, 104

fdes->inport, 30
fdes->outport, 30
fdes-flags, 32
fdes-status, 33
field-reader, 148
field-splitter, 144
file-directory?, 43
file-executable?, 44
file-exists?, 44
file-fifo?, 43
file-group, 42
file-info, 41
file-info:atime, 42
file-info:ctime, 42
file-info:device, 42
file-info:gid, 42
file-info:inode, 42
file-info:mode, 42
file-info:mtime, 42
file-info:nlinks, 42
file-info:size, 42
file-info:type, 42
file-info:uid, 42
file-inode, 42
file-last-access, 42
file-last-mod, 42
file-last-status-change, 42
file-match, 47
file-mode, 42
file-name-absolute?, 101
file-name-as-directory, 101
file-name-directory, 102
file-name-directory?, 101
file-name-extension, 103
file-name-non-directory?, 101
file-name-nondirectory, 102
file-name-sans-extension, 103
file-nlinks, 42
file-not-executable?, 43
file-not-exists?, 44

file-not-readable?, 43
file-not-writable?, 43
file-owner, 42
file-readable?, 44
file-regular?, 43
file-size, 42
file-socket?, 43
file-special?, 43
file-symlink?, 43
file-type, 42
file-writable?, 44
fill-in-date!, 72
flush-all-ports, 37
flush-submatches, 135
flush-tty/both, 82
flush-tty/input, 82
flush-tty/output, 82
force-output, 37
fork, 53
fork-pty-session, 83
fork/pipe, 54
fork/pipe+, 55
format-date, 72

get-lock-region, 39
getenv, 74
glob, 45
glob-quote, 47
group-info, 62
group-info, 62
group-info:gid, 62
group-info:members, 62
group-info:name, 62

home-dir, 105
home-directory, 77
home-file, 105
host-info, 97

if-match, 134
if-sre-form, 138
index, 105
infix-splitter, 144
internet-address->socket-address, 93
interrupt-handler, 68
interrupt-set, 67
itimer, 64

join-strings, 147

let-match, 134
listen-socket, 94
lock-region, 39
lock-region/no-block, 39
lock-region:end, 38
lock-region:exclusive?, 38
lock-region:len, 38
lock-region:proc, 38
lock-region:start, 38
lock-region:whence, 38
lock-region?, 38

make-date, 70
make-lock-region, 39
make-pty-generator, 84
make-re-char-set, 137
make-re-choice, 136
make-re-dsm, 137
make-re-repeat, 136
make-re-seq, 136
make-re-string, 137
make-re-submatch, 136
make-regexp, 129
make-string-input-port, 26
make-string-output-port, 26
make-tty-info, 80
match-cond, 134
match:end, 130
match:start, 130
match:substring, 130
move->fdes, 30

network-info, 97
nice, 60
nth, 156

open-control-tty, 82
open-fdes, 32
open-file, 32

open-input-file, 32
open-output-file, 32
open-pty, 84

parent-pid, 60
parse-file-name, 104
parse-sre, 139
parse-sres, 139
path-list->file-name, 103
pause-until-interrupt, 64
pid, 60
pid->proc, 55
pipe, 33
port->fdes, 30
port->list, 15
port->sexp-list, 15
port->string, 15
port->string-list, 15
port-fold, 16
port-revealed, 30
posix-string->regexp, 135
predicate->char-set, 107
priority, 60
proc, 55
proc:pid, 55
proc?, 55
process-group, 60
process-times, 61
protocol-info, 97
pty-name->tty-name, 84

re-any, 138
re-any?, 138
re-bol, 137
re-bol?, 137
re-bos, 137
re-bos?, 137
re-bow, 137
re-bow?, 137
re-char-set, 137
re-char-set:cset, 137
re-char-set?, 137
re-choice, 136
re-choice:elts, 136
re-choice:tsm, 136
re-choice?, 136
re-dsm, 137
re-dsm:body, 137
re-dsm:post-dsm, 137
re-dsm:pre-dsm, 137
re-dsm:tsm, 137
re-dsm?, 137
re-empty, 138
re-empty?, 138
re-eol, 137
re-eol?, 137
re-eos, 137
re-eos?, 137
re-eow, 137
re-eow?, 137
re-nonl, 138
re-repeat:from, 136
re-repeat:to, 136
re-repeat:tsm, 136
re-repeat?, 136
re-seq, 136
re-seq:elts, 136
re-seq:tsm, 136
re-seq?, 136
re-string, 137
re-string:chars, 137
re-string?, 137
re-submatch:post-dsm, 137
re-submatch:pre-dsm, 137
re-submatch:tsm, 137
re-submatch?, 136
re-trivial, 137
re-trivial?, 137
re-tsm, 138
re-word, 138
read-delimited, 141
read-delimited!, 141
read-line, 140
read-paragraph, 141
read-string, 33
read-string!, 33
read-string!/partial, 34
read-string/partial, 34

read-symlink, 40
reap-zombies, 57
receive-message, 95
receive-message!, 95
receive-message!/partial, 95
receive-message/partial, 95
record-reader, 144
regexp->posix-string, 135
regexp->scheme, 139
regexp->sre, 135
regexp-fold, 132
regexp-fold-right, 133
regexp-for-each, 134
regexp-search, 130
regexp-search?, 130
regexp-substitute, 130
regexp-substitute/global, 131
regexp?, 130, 138
release-port-handle, 30
rename-file, 40
repl, 156
replace-extension, 104
resolve-file-name, 104
rindex, 105
run, 13
run, 13
run/collecting, 17
run/collecting*, 17
run/file, 14
run/file*, 15
run/port, 14
run/port*, 15
run/port+proc, 17
run/port+proc*, 17
run/sexp, 14
run/sexp*, 15
run/sexps, 14
run/sexps*, 15
run/string, 14
run/string*, 15
run/strings, 14
run/strings*, 15
rx, 129
seek, 31
select , 34
send-message, 95
send-message/partial, 95
send-tty-break, 81
service-info, 97
set-enabled-interrupts, 67
set-fdes-flags, 32
set-fdes-status, 33
set-file-group, 41
set-file-mode, 41
set-file-owner, 41
set-file-times, 41
set-gid, 61
set-interrupt-handler, 67
set-port-buffering, 37
set-priority, 60
set-process-group, 60
set-socket-option, 96
set-tty-info/drain, 81
set-tty-info/flush, 81
set-tty-info/now, 81
set-tty-process-group, 83
set-uid, 61
set-umask, 60
setenv, 74
shutdown-socket, 95
signal->interrupt, 67
signal-process, 64
signal-process-group, 64
simplify-file-name, 104
simplify-regexp, 135
skip-char-set, 142
sleep, 64
sleep-until, 64
sloppy-suffix-splitter, 144
socket-address->internet-address, 93
socket-address->unix-address, 93
socket-connect, 90
socket-local-address, 94
socket-option, 96
socket-remote-address, 94

split-file-name, 102
sre->regexp, 135
sre-form?, 139
start-tty-input, 82
start-tty-output, 82
static-regexp?, 139
status:exit-val, 59
status:stop-sig, 59
status:term-sig, 59
stdio->stdports, 25
stdports->stdio, 25
stop-tty-input, 82
stop-tty-output, 82
string->char-set, 107
string-filter, 20
string-match, 129
string-output-port-output, 26
substitute-env-vars, 106
suffix-splitter, 144
suspend, 53
sync-file, 41
sync-file-system, 41
system-name, 64

tell, 31
temp-file-channel, 51
temp-file-iterate, 50
ticks/sec, 70
time, 69
time, 71
time+ticks, 70
truncate-file, 41
tty-file-name, 78
tty-info, 81
tty-info record type, 78
tty-info:control-chars, 78
tty-info:control-flags, 78
tty-info:input-flags, 78
tty-info:input-speed, 78
tty-info:local-flags, 78
tty-info:min, 78
tty-info:output-flags, 78
tty-info:output-speed, 78
tty-info:time, 78
tty-info?, 78
tty-name->pty-name, 84
tty-process-group, 83
tty?, 78
ttyc/2-stop-bits, 88
ttyc/carrier-flow-ctl, 88
ttyc/char-size, 88
ttyc/char-size5, 88
ttyc/char-size6, 88
ttyc/char-size7, 88
ttyc/char-size8, 88
ttyc/CTS-output-flow-ctl, 88
ttyc/enable-parity, 88
ttyc/enable-read, 88
ttyc/hup-on-close, 88
ttyc/ignore-flags, 88
ttyc/no-modem-sync, 88
ttyc/odd-parity, 88
ttyc/RTS-input-flow-ctl, 88
ttychar/delayed-suspend, 85
ttychar/delete-char, 85
ttychar/delete-line, 85
ttychar/delete-word, 85
ttychar/discard, 85
ttychar/eof, 85
ttychar/eol, 85
ttychar/eol2, 85
ttychar/interrupt, 85
ttychar/literal-next, 85
ttychar/quit, 85
ttychar/reprint, 85
ttychar/start, 85
ttychar/status, 85
ttychar/stop, 85
ttychar/suspend, 85
ttyin/7bits, 86
ttyin/beep-on-overflow, 86
ttyin/check-parity, 86
ttyin/cr->nl, 86
ttyin/ignore-bad-parity-chars, 86
ttyin/ignore-break, 86
ttyin/ignore-cr, 86
ttyin/input-flow-ctl, 86

ttyin/interrupt-on-break, 86
ttyin/lowercase, 86
ttyin/mark-parity-errors, 86
ttyin/nl->cr, 86
ttyin/output-flow-ctl, 86
ttyin/xon-any, 86
ttyl/alt-delete-word, 89
ttyl/canonical, 89
ttyl/case-map, 89
ttyl/echo, 89
ttyl/echo-ctl, 89
ttyl/echo-delete-line, 89
ttyl/echo-nl, 89
ttyl/enable-signals, 89
ttyl/extended, 89
ttyl/flush-output, 89
ttyl/hardcopy-delete, 89
ttyl/no-flush-on-interrupt, 89
ttyl/no-kernel-status, 89
ttyl/reprint-unread-chars, 89
ttyl/ttou-signal, 89
ttyl/visual-delete, 89
ttyl/visual-delete-line, 89
ttyout/all-delay, 87
ttyout/bs-delay, 87
ttyout/bs-delay0, 87
ttyout/bs-delay1, 87
ttyout/cr->nl, 86
ttyout/cr-delay, 87
ttyout/cr-delay0, 87
ttyout/cr-delay1, 87
ttyout/cr-delay2, 87
ttyout/cr-delay3, 87
ttyout/delay-w/fill-char, 86
ttyout/discard-eot, 86
ttyout/enable, 86
ttyout/expand-tabs, 86
ttyout/ff-delay, 87
ttyout/ff-delay0, 87
ttyout/ff-delay1, 87
ttyout/fill-w/del, 86
ttyout/nl->crnl, 86
ttyout/nl-delay, 87
ttyout/nl-delay0, 87
ttyout/nl-delay1, 87
ttyout/nl-does-cr, 86
ttyout/no-col0-cr, 86
ttyout/tab-delay, 87
ttyout/tab-delay0, 87
ttyout/tab-delay1, 87
ttyout/tab-delay2, 87
ttyout/tab-delayx, 87
ttyout/uppercase, 86
ttyout/vtab-delay, 87
ttyout/vtab-delay0, 87
ttyout/vtab-delay1, 87

umask, 60
uncase, 135
uncase-char-set, 135
uncase-string, 135
unix-address->socket-address, 93
unlock-region, 39
user-effective-gid, 61
user-effective-uid, 61
user-gid, 61
user-info, 61
user-info, 61
user-info:gid, 61
user-info:home-dir, 61
user-info:name, 61
user-info:shell, 61
user-info:uid, 61
user-login-name, 61
user-supplementary-gids, 61
user-uid, 61

wait, 58
wait-any, 58
wait-process-group, 58
with-current-input-port, 24
with-current-input-port*, 24
with-current-output-port, 24
with-current-output-port*, 24
with-cwd, 60
with-cwd*, 60
with-enabled-interrupts, 67

with-enabled-interrupts*, 67
with-env, 75
with-env*, 75
with-errno-handler, 22
with-errno-handler*, 22
with-error-output-port, 24
with-error-output-port*, 24
with-region-lock, 39
with-region-lock*, 39
with-stdio-ports, 26
with-stdio-ports*, 26
with-total-env, 75
with-total-env*, 75
with-umask, 60
with-umask*, 60
write-string, 35
write-string/partial, 36