Glossary

Action

A series of *awk* statements attached to a rule. If the rule's pattern matches an input record, *awk* executes the rule's action. Actions are always enclosed in curly braces. (See the section "Actions" in Chapter 6, *Patterns, Actions, and Variables*.)

Amazing awk Assembler

Henry Spencer at the University of Toronto wrote a retargetable assembler completely as *sed* and *awk* scripts. It is thousands of lines long, including machine descriptions for several eight-bit microcomputers. It is a good example of a program that would have been better written in another language. It is available over the Internet from *ftp://ftp.freefriends.org/arnold/Awkstuff/aaa.tgz*.

Amazingly Workable Formatter (awf)

Henry Spencer at the University of Toronto wrote a formatter that accepts a large subset of the `nroff -ms` and `nroff -man` formatting commands, using *awk* and *sh*. It is available over the Internet from *ftp://ftp.freefriends.org/arnold/Awkstuff/awf.tgz*.

Anchor

The regexp metacharacters `^` and `$`, which force the match to the beginning or end of the string, respectively.

ANSI

The American National Standards Institute. This organization produces many standards, among them the standards for the C and C++ programming languages. These standards often become international standards as well. See also "ISO."

Array
: A grouping of multiple values under the same name. Most languages provide just sequential arrays. *awk* provides associative arrays (see "Associative Array").

Assertion
: A statement in a program that a condition is true at this point in the program. Useful for reasoning about how a program is supposed to behave.

Assignment
: An *awk* expression that changes the value of some *awk* variable or data object. An object that you can assign to is called an *lvalue*. The assigned values are called *rvalues*. See the section "Assignment Expressions" in Chapter 5, *Expressions*.

Associative Array
: Arrays in which the indices may be numbers or strings, not just sequential integers in a fixed range.

awk Language
: The language in which *awk* programs are written.

awk Program
: An *awk* program consists of a series of *patterns* and *actions*, collectively known as *rules*. For each input record given to the program, the program's rules are all processed in turn. *awk* programs may also contain function definitions.

awk Script
: Another name for an *awk* program.

Bash
: The GNU version of the standard shell (the Bourne-again shell). See also "Bourne Shell."

BBS
: See "Bulletin Board System."

Bit Short for "Binary Digit." All values in computer memory ultimately reduce to binary digits: values that are either zero or one. Groups of bits may be interpreted differently—as integers, floating-point numbers, character data, addresses of other memory objects, or other data. *awk* lets you work with floating-point numbers and strings. *gawk* lets you manipulate bit values with the built-in functions described in the section "Bit-Manipulation Functions of gawk" in Chapter 8, *Functions*.

Computers are often defined by how many bits they use to represent integer values. Typical systems are 32-bit systems, but 64-bit systems are becoming increasingly popular, while 16-bit systems are waning in popularity.

Boolean Expression
: Named after the English mathematician Boole. See also "Logical Expression."

Bourne Shell
: The standard shell (*/bin/sh*) on Unix and Unix-like systems, originally written by Steven R. Bourne. Many shells (Bash, *ksh*, *pdksh*, *zsh*) are generally upwardly compatible with the Bourne shell.

Built-in Function
: The *awk* language provides built-in functions that perform various numerical, I/O-related, and string computations. Examples are `sqrt` (for the square root of a number) and `substr` (for a substring of a string). *gawk* provides functions for timestamp management, bit manipulation, and runtime string translation. (See the section "Built-in Functions" in Chapter 8.)

Built-in Variable
: `ARGC`, `ARGV`, `CONVFMT`, `ENVIRON`, `FILENAME`, `FNR`, `FS`, `NF`, `NR`, `OFMT`, `OFS`, `ORS`, `RLENGTH`, `RSTART`, `RS`, and `SUBSEP` are the variables that have special meaning to *awk*. In addition, `ARGIND`, `BINMODE`, `ERRNO`, `FIELDWIDTHS`, `IGNORECASE`, `LINT`, `PROCINFO`, `RT`, and `TEXTDOMAIN` are the variables that have special meaning to *gawk*. Changing some of them affects *awk*'s running environment. (See the section "Built-in Variables" in Chapter 6.)

Braces
: See "Curly Braces."

Bulletin Board System
: A computer system allowing users to log in and read and/or leave messages for other users of the system, much like leaving paper notes on a bulletin board.

C
: The system programming language that most GNU software is written in. The *awk* programming language has C-like syntax, and this book points out similarities between *awk* and C when appropriate.

  In general, *gawk* attempts to be as similar to the 1990 version of ISO C as makes sense. Future versions of *gawk* may adopt features from the newer 1999 standard, as appropriate.

C++
: A popular object-oriented programming language derived from C.

Character Set
: The set of numeric codes used by a computer system to represent the characters (letters, numbers, punctuation, etc.) of a particular country or place. The most common character set in use today is ASCII (American Standard Code for Information Interchange). Many European countries use an extension of ASCII known as ISO-8859-1 (ISO Latin-1).

CHEM
: A preprocessor for *pic* that reads descriptions of molecules and produces *pic* input for drawing them. It was written in *awk* by Brian Kernighan and Jon Bentley, and is available from *http://cm.bell-labs.com/netlib/typesetting/chem.gz*.

Coprocess
: A subordinate program with which two-way communication is possible.

Compiler
: A program that translates human-readable source code into machine-executable object code. The object code is then executed directly by the computer. See also "Interpreter."

Compound Statement
: A series of *awk* statements, enclosed in curly braces. Compound statements may be nested. (See the section "Control Statements in Actions" in Chapter 6.)

Concatenation
: Concatenating two strings means sticking them together, one after another, producing a new string. For example, the string `foo` concatenated with the string `bar` gives the string `foobar`. (See the section "String Concatenation" in Chapter 5.)

Conditional Expression
: An expression using the `?:` ternary operator, such as *expr1* `?` *expr2* `:` *expr3*. The expression *expr1* is evaluated; if the result is true, the value of the whole expression is the value of *expr2*; otherwise, the value is *expr3*. In either case, only one of *expr2* and *expr3* is evaluated. (See the section "Conditional Expressions" in Chapter 5.)

Comparison Expression
: A relation that is either true or false, such as `(a < b)`. Comparison expressions are used in `if`, `while`, `do`, and `for` statements, and in patterns to select which input records to process. (See the section "Variable Typing and Comparison Expressions" in Chapter 5.)

Curly Braces
: The characters `{` and `}`. Curly braces are used in *awk* for delimiting actions, compound statements, and function bodies.

Dark Corner
: An area in the language in which specifications often were (or still are) not clear, leading to unexpected or undesirable behavior. Such areas are marked with "(d.c.)" in the text and are indexed under the heading "dark corner."

Data Driven
: A description of *awk* programs, in which you specify the data you are interested in processing and what to do when that data is seen.

Data Objects
: Numbers and strings of characters. Numbers are converted into strings and vice versa, as needed. (See the section "Conversion of Strings and Numbers" in Chapter 5.)

Deadlock
: The situation in which two communicating processes are each waiting for the other to perform an action.

Double-Precision
: An internal representation of numbers that can have fractional parts. Double-precision numbers keep track of more digits than do single-precision numbers, but operations on them are sometimes more expensive. This is the way *awk* stores numeric values. It is the C type `double`.

Dynamic Regular Expression
: A dynamic regular expression is a regular expression written as an ordinary expression. It could be a string constant, such as `"foo"`, but it may also be an expression whose value can vary. (See the section "Using Dynamic Regexps" in Chapter 2, *Regular Expressions*.)

Environment
: A collection of strings, of the form *name=val*, that each program has available to it. Users generally place values into the environment in order to provide information to various programs. Typical examples are the environment variables HOME and PATH.

Empty String
: See "Null String."

Epoch
: The date used as the "beginning of time" for timestamps. Time values in Unix systems are represented as seconds since the epoch, with library functions available for converting these values into standard date and time formats.

  The epoch on Unix and POSIX systems is 1970-01-01 00:00:00 UTC. See also "GMT" and "UTC."

Escape Sequences
: A special sequence of characters used for describing nonprinting characters, such as `\n` for newline or `\033` for the ASCII ESC (Escape) character. (See the section "Escape Sequences" in Chapter 2.)

FDL
: See "Free Documentation License."

Field
: When *awk* reads an input record, it splits the record into pieces separated by whitespace (or by a separator regexp that you can change by setting the built-in variable `FS`). Such pieces are called fields. If the pieces are of fixed length, you can use the built-in variable `FIELDWIDTHS` to describe their lengths. (See the section "Specifying How Fields Are Separated" and the section "Reading Fixed-Width Data" in Chapter 3, *Reading Input Files*.)

Flag
: A variable whose truth value indicates the existence or nonexistence of some condition.

Floating-Point Number
: Often referred to in mathematical terms as a "rational" or real number, this is just a number that can have a fractional part. See also "Double-Precision" and "Single-Precision."

Format
: Format strings are used to control the appearance of output in the `strftime` and `sprintf` functions, and are used in the `printf` statement as well. Also, data conversions from numbers to strings are controlled by the format string contained in the built-in variable `CONVFMT`. (See the section "Format-Control Letters" in Chapter 4, *Printing Output*.)

Free Documentation License
: This document describes the terms under which this book is published and may be copied. (See Appendix F, *GNU Free Documentation License*.)

Function
: A specialized group of statements used to encapsulate general or program-specific tasks. *awk* has a number of built-in functions, and also allows you to define your own. (See Chapter 8.)

FSF
: See "Free Software Foundation."

Free Software Foundation
: A nonprofit organization dedicated to the production and distribution of freely distributable software. It was founded by Richard M. Stallman, the author of the original Emacs editor. GNU Emacs is the most widely used version of Emacs today.

gawk
: The GNU implementation of *awk*.

General Public License
: This document describes the terms under which *gawk* and its source code may be distributed. (See Appendix E, *GNU General Public License*.)

GMT
: "Greenwich Mean Time." This is the old term for UTC. It is the time of day used as the epoch for Unix and POSIX systems. See also "Epoch" and "UTC."

GNU
: "GNU's not Unix." An ongoing project of the Free Software Foundation to create a complete, freely distributable, POSIX-compliant computing environment.

GNU/Linux
: A variant of the GNU system using the Linux kernel, instead of the Free Software Foundation's Hurd kernel. Linux is a stable, efficient, full-featured clone of Unix that has been ported to a variety of architectures. It is most popular on PC-class systems, but runs well on a variety of other systems too. The Linux kernel source code is available under the terms of the GNU General Public License, which is perhaps its most important aspect.

GPL
: See "General Public License."

Hexadecimal
: Base 16 notation, in which the digits are `0`–`9` and `A`–`F`, with `A` representing 10, `B` representing 11, and so on, up to `F` for 15. Hexadecimal numbers are written in C using a leading `0x`, to indicate their base. Thus, `0x12` is 18 (1 times 16 plus 2).

I/O
: Abbreviation for "input/output," the act of moving data into and/or out of a running program.

Input Record
: A single chunk of data that is read in by *awk*. Usually, an *awk* input record consists of one line of text. (See the section "How Input Is Split into Records" in Chapter 3.)

Integer
: A whole number, i.e., a number that does not have a fractional part.

Internationalization
: The process of writing or modifying a program so that it can use multiple languages without requiring further source code changes.

Interpreter
: A program that reads human-readable source code directly, and uses the instructions in it to process data and produce results. *awk* is typically (but not always) implemented as an interpreter. See also "Compiler."

Interval Expression
: A component of a regular expression that lets you specify repeated matches of some part of the regexp. Interval expressions were not traditionally available in *awk* programs.

ISO
: The International Standards Organization. This organization produces international standards for many things, including programming languages, such as C and C++. In the computer arena, important standards like those for C, C++, and POSIX become both American national and ISO international standards simultaneously. This book refers to Standard C as "ISO C" throughout.

Keyword
: In the *awk* language, a keyword is a word that has special meaning. Keywords are reserved and may not be used as variable names.

  gawk's keywords are: `BEGIN`, `END`, `if`, `else`, `while`, `do...while`, `for`, `for...in`, `break`, `continue`, `delete`, `next`, `nextfile`, `function`, `func`, and `exit`.

Lesser General Public License (LGPL)
: This document describes the terms under which binary library archives or shared objects, and whether their source code may be distributed.

Linux
: See "GNU/Linux."

Localization
: The process of providing the data necessary for an internationalized program to work in a particular language.

Logical Expression
: An expression using the operators for logic, AND, OR, and NOT, written `&&`, `||`, and `!` in *awk*. Often called Boolean expressions, after the mathematician who pioneered this kind of mathematical logic.

Lvalue
: An expression that can appear on the left side of an assignment operator. In most languages, *lvalues* can be variables or array elements. In *awk*, a field designator can also be used as an *lvalue*.

Matching
: The act of testing a string against a regular expression. If the regexp describes the contents of the string, it is said to *match* it.

Metacharacters
: Characters used within a regexp that do not stand for themselves. Instead, they denote regular expression operations, such as repetition, grouping, or alternation.

Null String
: A string with no characters in it. It is represented explicitly in *awk* programs by placing two double quote characters next to each other (`""`). It can appear in input data by having two successive occurrences of the field separator appear next to each other.

Number
: A numeric-valued data object. Modern *awk* implementations use double-precision floating-point to represent numbers. Very old *awk* implementations use single-precision floating-point.

Octal
: Base-8 notation, in which the digits are `0`–`7`. Octal numbers are written in C using a leading `0`, to indicate their base. Thus, `013` is 11 (1 times 8 plus 3).

P1003.2
: See "POSIX."

Pattern
: Patterns tell *awk* which input records are interesting to which rules.

  A pattern is an arbitrary conditional expression against which input is tested. If the condition is satisfied, the pattern is said to *match* the input record. A typical pattern might compare the input record against a regular expression. (See the section "Pattern Elements" in Chapter 6.)

POSIX
: The name for a series of standards that specify a Portable Operating System interface. The "IX" denotes the Unix heritage of these standards. The main standard of interest for *awk* users is *IEEE Standard for Information Technology, Standard 1003.2-1992, Portable Operating System Interface (POSIX) Part 2: Shell and Utilities*. Informally, this standard is often referred to as simply "P1003.2."

Precedence
: The order in which operations are performed when operators are used without explicit parentheses.

Private
: Variables and/or functions that are meant for use exclusively by library functions and not for the main *awk* program. Special care must be taken when naming such variables and functions. (See the section "Naming Library Function Global Variables" in Chapter 12, *A Library of awk Functions*.)

Range (of input lines)
: A sequence of consecutive lines from the input file(s). A pattern can specify ranges of input lines for *awk* to process or it can specify single lines. (See the section "Pattern Elements" in Chapter 6.)

Recursion
: When a function calls itself, either directly or indirectly. If this isn't clear, refer to the entry for "redirection."

Redirection
: Redirection means performing input from something other than the standard input stream, or performing output to something other than the standard output stream.

  You can redirect the output of the `print` and `printf` statements to a file or a system command, using the `>`, `>>`, `|`, and `|&` operators. You can redirect input to the `getline` statement using the `<`, `|`, and `|&` operators. (See the section "Redirecting Output of print and printf" in Chapter 4, and the section "Explicit Input with getline" in Chapter 3.)

Regexp
: Short for *regular expression*. A regexp is a pattern that denotes a set of strings, possibly an infinite set. For example, the regexp `R.*xp` matches any string starting with the letter `R` and ending with the letters `xp`. In *awk*, regexps are used in patterns and in conditional expressions. Regexps may contain escape sequences. (See Chapter 2.)

Regular Expression
: See "Regexp."

Regular Expression Constant
: A regular expression constant is a regular expression written within slashes, such as `/foo/`. This regular expression is chosen when you write the *awk* program and cannot be changed during its execution. (See the section "How to Use Regular Expressions" in Chapter 2.)

Rule
: A segment of an *awk* program that specifies how to process single-input records. A rule consists of a *pattern* and an *action*. *awk* reads an input record; then, for each rule, if the input record satisfies the rule's pattern, *awk* executes the rule's action. Otherwise, the rule does nothing for that input record.

Rvalue
: A value that can appear on the right side of an assignment operator. In *awk*, essentially every expression has a value. These values are *rvalues*.

Scalar
: A single value, be it a number or a string. Regular variables are scalars; arrays and functions are not.

Search Path
: In *gawk*, a list of directories to search for *awk* program source files. In the shell, a list of directories to search for executable programs.

Seed
: The initial value, or starting point, for a sequence of random numbers.

sed
: See "Stream Editor."

Shell
: The command interpreter for Unix and POSIX-compliant systems. The shell works both interactively, and as a programming language for batch files or shell scripts.

Short-Circuit
: The nature of the *awk* logical operators `&&` and `||`. If the value of the entire expression is determinable from evaluating just the lefthand side of these operators, the righthand side is not evaluated. (See the section "Boolean Expressions" in Chapter 5.)

Side Effect
: A side effect occurs when an expression has an effect aside from merely producing a value. Assignment expressions, increment and decrement expressions, and function calls have side effects. (See the section "Assignment Expressions" in Chapter 5.)

Single-Precision
: An internal representation of numbers that can have fractional parts. Single-precision numbers keep track of fewer digits than do double-precision numbers, but operations on them are sometimes less expensive in terms of CPU time. This is the type used by some very old versions of *awk* to store numeric values. It is the C type `float`.

Space
: The character generated by hitting the space bar on the keyboard.

Special File
: A filename interpreted internally by *gawk*, instead of being handed directly to the underlying operating system—for example, */dev/stderr*. (See the section "Special Filenames in gawk" in Chapter 4.)

Stream Editor
: A program that reads records from an input stream and processes them one or more at a time. This is in contrast with batch programs, which may expect to read their input files in entirety before starting to do anything, as well as with interactive programs that require input from the user.

String
: A datum consisting of a sequence of characters, such as `I am a string`. Constant strings are written with double quotes in the *awk* language and may contain escape sequences. (See the section "Escape Sequences" in Chapter 2.)

Tab
: The character generated by hitting the Tab key on the keyboard. It usually expands to up to eight spaces upon output.

Text Domain
: A unique name that identifies an application. Used for grouping messages that are translated at runtime into the local language.

Timestamp
: A value in the "seconds since the epoch" format used by Unix and POSIX systems. Used for the *gawk* functions `mktime`, `strftime`, and `systime`. See also "Epoch" and "UTC."

Unix
: A computer operating system originally developed in the early 1970s at AT&T Bell Laboratories. It initially became popular in universities around the world and later moved into commercial environments as a software-development system and network-server system. There are many commercial versions of Unix, as well as several work-alike systems whose source code is freely available (such as GNU/Linux, NetBSD, FreeBSD, and OpenBSD).

UTC
: The accepted abbreviation for "Universal Coordinated Time." This is standard time in Greenwich, England, which is used as a reference time for day and date calculations. See also "Epoch" and "GMT."

Whitespace
: A sequence of space, tab, or newline characters occurring inside an input record or a string.