

Regular Expressions

A **regular expression** is a template or pattern to be matched against a string.

In Perl a regular expression is enclose inside two slashes:

/regular_expression/

The regular expression may contain:

- Ordinary text to be matched to an exact pattern (or sub pattern)
- Special operator characters characters that have a special meaning and control how we match patterns







Alternative Character Matches

Square brackets ([..]) allow for any one of the letters listed inside the brackets to be matched at the specified position.

For example /b[aiu]t/ can only match to **bat**, **bit** or **but**.

You can specify a **range of values** inside [..].

For example:

```
[012345679] # any single digit
[0-9] # also any single digit
[a-z] # any single lower case letter
[a-zA-Z] # any single letter
[0-9\-] # 0-9 plus minus character
```















```
CARDIF
  Parenthesis ( . . ) For Enforcing Precedence
                                                                         PRIFYSGOL
  Parenthesis ( ... ) can be used to delimit special matches and
therefore enforce precedence.
                                                                         Internet
Computing
CM0133
 • Much like you control evaluation of arithmetic and other
   expressions in any programming language
                                                                          546
  For example:
(abc) *
  matches " ", abc, abcabc, abcabcabc, .....
  and
                                                                           (a|b) (c|d)
                                                                           ••
  matches ac, ad, bc, bd
                                                                           ◀
                                                                          Back
                                                                          Close
```







Memory Variables Example

```
$_ = "One Two Three Four Once ....";
/(\w+)\W+(\w+)/; # match first two words
```

print "1st Word is " . \$1" . "\n";
print "2nd Word is " . \$2" . "\n";

How does this work?

- We check for valid sequences of characters Not Valid English Words
- \w matches any word character
- w+ matches one or more word characters
- (\w+) remembers respective matches
- \W matches any **Non**-word character
- \W+ matches one or more **Non**-word characters
- \W+ delimits any break between two sets of word characters
- \$1 and \$2 loaded with two sets of respective word characters.



Suppressing Memories

(?:regular_expression)

• This groups things like regular () but doesn't make back references to internal or external memories like () does.







Substitution

We may frequently need to change a pattern in a string.

The **substitution operator** (s) is the simplest form of substitution. It has the form:

s/old_regex/new_string/

Note: We can (optionally) qualify the substitution with

- A g global substitution,
- A i ignore case and
- A e evaluate right side as expression and others.

We place the qualifiers, if appropriate, on the right hand side of the substitution, For example:

s/old_regex/new_string/gie

Zero, one or more qualifiers may be used.



CARDIF|

Practical CGI Related Example

To replace the + characters from CGI input with a space we could do:

\$CGI_in_val = s/\+/ /ge;

- We use g global substitution to replace all occurrences of + in the input string.
- We need e to force the evaluations of the expression so that the value can be returned to the input string:
 - That is to say the input string will have its value changed, for future use in the Perl code, by this operation

Note: cgi-lib.pl actually does this and other cgi character conversion automatically — peek at the cgi-lib.pl source code and find such examples



CARDIF| Split() and join() UNIVERSI AERDY Split () and join () are two very useful functions. Split () takes a regular expression and a string: Internet Computing CM0133 split(reg_ex, string) 556 and looks for all occurrences of the regular expression and the parts of the string that don't match are returned, in sequence, in a list. Example: To split an input name/value pair CGI input we could do: \$cgi_pair = "name=value"; #format of input (\$name,\$value) = split(/=/,\$cgi_pair); 44 @cgi_list = = split(/=/,\$cgi_pair); •• The join() function takes two lists and glues them together. Back Close



UNIX Read Directory Example

```
On UNIX we may do:
```

```
opendir(IDIR,"./Internet")
   || die "NO SUCH Directory: Images";
while ($file = readdir(DIR) )
   {
        print " $file\n";
    }
closedir(DIR);
```

The above reads a sub-directory Internet assumed to be located in the same directory from where the Perl script has been run ./.

Note: The || die "..." is a short hand if type statement that outputs a string, enclosed in "..." and then quits the program





Reading and Writing Files

We have just introduced the concept of a **Directory Handle** for referring to a Directory on disk.

We now introduce a similar concept of **File Handle** for referring to a File on disk from which we can read data and to which we can write data.

Similar ideas of opening and closing the files also exist.

You use the open () operator to open a file (for reading):

open(FILEHANDLE, "file_on_device");

The file may be accessed with an absolute or relative path.









Writing to a File

To write to a file you use the print command and simply refer to the FILEHANDLE before you format and output the string, I.e.:

```
print FILEHANDLE "Output String\n";
```

Example: To read from one file infile and copy line by line to another outfile we could do:

```
open(IN, "infile")
    || die "cannot open input file";
open(OUT, "outfile")
    || die "cannot open output file";
while(<IN>)
{ print OUT $_; # echo line read
}
close(IN);
close(OUT);
```









Some Example Perl Scripts

Let us conclude our brief study of Perl by looking at some example CGI scripts.

The examples are taken form the Laura Lemay book:

"Teach yourself web publishing in HTML" Series of books.



An Address Book Search Engine

This is a more complex and larger script:

- It illustrates how information may be queried from a information stored in a type database — for now we keep things simple.
- The data base is just a text file and we can only read from the file.

	w w w Address Manager
	Enter search values in any field.
	Name:
	Address:
	Home Phone:
	Work Phone:
	Email Address:
	Home Page:
	(Search) (Clear
To see this form in action <u>click here</u> .	

W Address Manager

CARDIF

PRIEYSGO CAERDY

CARDIFF UNIVERSITY Address Book Example — The HTML/Browser/Client Side PRIFYSGO CAERDY The HTML Form Front-End is composed via: <HTML> <HEAD> <TITLE>Address Book Search Forms</TITLE> Internet Computing </HEAD> CM0133 <BODY> <H1>WWW Address Manager</H1> 570 <P>Enter search values in any field. <PRE><HR> <FORM METHOD=POST ACTION="http://www.cs.cf.ac.uk/user/Dave.Marshall/cgi-bin/address.pl"> <P>Name: <INPUT TYPE="text" NAME="Name" SIZE=40> <P>Address: <INPUT TYPE="text" NAME="Address" SIZE=40> <P>Home Phone: <INPUT TYPE="text" NAME="Hphone" SIZE=40> <P>Work Phone: <INPUT TYPE="text" NAME="Wphone" SIZE=40> <P>Email Address: <INPUT TYPE="text" NAME="Email" SIZE=40> <P>Home Page: <INPUT TYPE="text" NAME="WWW" SIZE=40> </PRE> <INPUT TYPE="submit" VALUE="Search"> <INPUT TYPE="reset" VALUE="Clear"> •• <HR> ◀ </FORM> </BODY> Þ </HTML> Back Close
CARDIFF UNIVERSITY Address Book Example — The Perl/CGI/Server Side PRIFYSGO CAERDY The Perl CGI script is as follows: require 'cgi-lib.pl'; Internet # grab values passed from form: Computing &ReadParse(*in); CM0133 print "Content-type: text/html\n\n"; 571 # print the top part of the response print "<HTML><HEAD><TITLE>Addresss Book Search Results</TITLE></HEAD>\n"; print "<BODY><H1>Addresss Book Search Results</H1>\n"; # read and parse data file \$data="address.data"; open(DATA, "\$data") || die "Can't open \$data: \$!\n</BODY></HTML>\n"; while(<DATA>) { chop; # delete trailing \n if (/^\s*\$/) { # break between records if (\$match) { # if anything matched, print the whole record &printrecord(\$record); \$nrecords_matched++; •• } undef \$match; •• undef \$record; next; ◀ } # tag: value Þ (\$tag,\$val) = split(/:/,\$_,2); if (\$tag = ~ / Name/i) { Back Close

```
$match++ if( $in{'Name'} && $val = /\b$in{'Name'}\b/i) ;
$record = $val;
next;
}
if ($tag = / Address/i) {
$match++ if( $in{'Address'} && $val = /\b$in{'Address'}\b/i);
$record .= "\n<BR>$val" if ($val);
next;
if ($tag = / Home\s*Pho/i) {
$match++ if( $in{'Hphone'} && $val = '/b$in{'Hphone'}\b/i);
$record .= "\n<BR>Home: $val" if ($val);
next;
}
if ($tag = ~ / ^Work/i) {
$match++ if( $in{'Wphone'} && $val = /\b$in{'Wphone'}\b/i);
$record .= "\n<BR>Work: $val" if ($val);
next;
}
if ($tag = / ^Email/i) {
$match++ if( $in{'Email'} && $val = '\b$in{'Email'}\b/i) ;
$record .= "\n<BR><A HREF=\"mailto:$val\">$val</A>" if ($val);
next;
}
if ($tag = ~ /Page/i) {
$match++ if( $in{'WWW'} && $val = '/b$in{'WWW'}b/i);
$record .= "\n<BR><A HREF=$val>$val</A>" if ($val);
next;
}
# anything else
$record .= $_;
}
close DATA;
if (! defined $nrecords_matched)
{ print "<H2>No Matches</H2>\n"; }
```



```
CARDIFF
UNIVERSITY
PRIFYSGOL
CAERDYD
print "</BODY></HTML>\n";
exit;
sub printrecord {
local($buf) = @_;
print "<P>\n$buf<P>\n";
}
                                                                                                                                                 Internet
Computing
CM0133
                                                                                                                                                    573
                                                                                                                                                     ••
                                                                                                                                                     •
                                                                                                                                                      ◀
                                                                                                                                                      Back
                                                                                                                                                   Close
```

What is going on here?

This Perl Script essentially does the following:

- We use the cgi-lib.pl ReadParse subroutine to read the CGI input.
- We extract out the associated name value pairs.
- The data is read in from a file address.data.txt
- The data is searched using Perl regular expressions for given Names, Addresses etc. and matches stored and printed out.
- The subroutine printrecord prints out the matched record.



Creating a Guest Book	CARDIFF
The Guest book is a more complicated example:	CAERDYP
• An extension of our the address book example.	
• In a guestbook readers can post comments about you WWW pages.	Internet Computing CM0133
• We write to a file in this example	575
Guestbooks are quite common on WWW sites. The HTML Form is as follows: Not a response: Name: Email address: Tex:	
	••
	>>
POST CLEAR	
To see this form in action click here.	Back
	Close

CARDIFF UNIVERSITY GuestBook Example — The HTML/Browser/Client Side PRIFYSGO CAERDY The HTML Form Front-End is composed via: <HTML> <HEAD> Internet Computing <TITLE>Comments!</TITLE> CM0133 </HEAD> 576 </BODY> <!--GUESTBOOK--> <H1>Comments!</H2> <P>Here are comments people have left about my pages. Post your own using the form at the end of the page. <P>Comments list started on <!--STARTDATE--> Apr 4 2003 Last post on <!--LASTDATE--> Thu Aug 24 09:25:46 PDT 2003 44 <HR>Susan M. sus@monitor.com •• ◀ Tue Apr 10 05:57:09 EDT 2003 <P>This is the worst home page I have ever seen on the net. Please Back Close



Text:
 <TEXTAREA ROWS=15 COLS=60 NAME="body"> </TEXTAREA>
 <INPUT TYPE=submit VALUE="POST"> <INPUT TYPE=reset VALUE="CLEAR"> </FORM> <HR> </BODY> </HTML>



CARDIFF UNIVERSITY GuestBook Example — The Perl/CGI/Server Side PRIFYSGO CAERDY The Perl CGI is as follows: require 'cgi-lib.pl'; # grab values passed from form: Internet &ReadParse(*in); Computing CM0133 print "Content-type: text/html\n\n"; 579 # print the top part of the response print "<HTML><HEAD>\n"; print "<TITLE>Post Results</TITLE>\n"; print "</HEAD><BODY>\n"; # change to your favorite date format: \$date = `date`; chop(\$date); # trim \n # Grab the HTML file and make a file name for the temp file. \$file = "\$ENV{'PWD'}" . "\$ENV{'PATH_INFO'}"; \$tmp = \$file . ".tmp"; $t= s/{//{g}} = s/{//{g}}$ # make a unique tmp file name from the path \$tmp = "/tmp/\$tmp"; # if any fields are blank, then skip the post and inform user: if (!\$in{'name'} || !\$in{'address'} || !\$in{'body'}) { •• &err("You haven't filled in all the fields. Back up and try again."); } •• # reformat the body of the post. we want to preserve paragraph breaks. ◀ \$text = \$in{'body'}; \$text = s/\n/r/\n/g; \$text = s/\r/<P>/g; Þ \$text = s/\n\n/<P>/g; Back Close

```
CARDIFF
UNIVERSITY
t = s/n//g;
$text = s/<P><P>/<P>/g;
                                                                                                  PRIFYSGO
                                                                                                  CAERDY
# get an exclusive open on the tmp file, so
# two posts at the same time don't clobber each other.
for($count = 0; -f "$tmp"; $count++) {
# oh no. someone else is trying to update the message file. so we wait.
sleep(1);
                                                                                                 Internet
&err("Tmp file in use, giving up!") if ($count > 4); # but not for long
                                                                                                 Computing
                                                                                                 CM0133
}
open(TMP, ">$tmp") || &err("Can't open tmp file.");
                                                                                                   580
# open the HTML file
open(FILE, "<$file") ||</pre>
&err("Can't open file $file: $!");
# an HTMLBBS file. look through it for the HTML comments
# that denote stuff we want to change:
while(<FILE>) {
if (/<!--LASTDATE-->/) { print TMP "<!--LASTDATE--> $date \n"; }
elsif (/<!--GUESTBOOK-->/) {
print TMP "<!--GUESTBOOK-->\n";
$guestbook++;
}
elsif (/<!--POINTER-->/) {
# add this post
print TMP "<HR>";
print TMP "<B>$in{'name'} \n";
print TMP " <A HREF=mailto:$in{'address'}>
                   $in{'address'}</A></B> $date\n";
                                                                                                   ••
print TMP "<P> $text\n<!--POINTER-->\n";
                                                                                                    ••
else { print TMP $_; } # copy lines
                                                                                                    ◀
if (! defined $guestbook)
                                                                                                    Þ
{ &err("not a Guestbook file!"); }
                                                                                                  Back
```

Close





CARDIFF UNIVERSITY A Web Page Counter PRIFYSGOL Web page counters can be used to indicate how many time your Web page or pages have been visited. Internet Computing CM0133 The URL: 583 http://www.htmlgoodies.com/beyond/countcgi.html contains Perl scripts, instructions and explanations as to how to achieve this. • ◀ ۲ Back Close