mathopd 4241 mathopd.conf 4244 mathopd.pid 4242 rundig 4317 w3m 4239 webalizer 4331 webalizer.conf 4331 wget 4342 wgetrc 4344

Il modo più comune per diffondere informazioni attraverso la rete è quello di utilizzare un servente HTTP (*Hypertext transfer protocol*). Le informazioni pubblicate in questo modo sono rivolte a tutti gli utenti che possono raggiungere il servizio, nel senso che normalmente non viene richiesta alcuna identificazione: al massimo si impedisce o si concede l'accesso in base al meccanismo di filtro gestito dal supervisore dei servizi di rete o dal TCP wrapper.

Per offrire un servizio HTTP occorre un programma in grado di gestirlo, di solito in forma di demone. Analogamente al servizio FTP anonimo, il servente HTTP consente l'accesso a una directory particolare e alle sue discendenti; questa directory viene identificata spesso con il nome *document root*. Quando il servente HTTP è in grado di distinguere con quale nome a dominio è stato raggiunto il servizio e, in base a tale nome, offre l'accesso a una directory differente, si dice che distingue tra i ***domini virtuali***.

Un servente HTTP non offre solo un servizio di semplice consultazione di documenti: permette anche di interpellare dei programmi. Questi programmi sono collocati normalmente al di fuori della directory da cui si diramano i documenti (HTML o di altro tipo), per evitare che questi possano essere letti. In questo contesto, tali programmi sono definiti ***gateway*** e normalmente vengono chiamati ***programmi CGI***, o *cgi-bin*. Ma l'avvio di programmi implica l'attribuzione di privilegi: di solito si fa in modo che questi funzionino utilizzando la personalità di un utente fittizio apposito ('**www**', '**nobody**' o simile), per evitare che possano compiere più azioni del necessario.

Secondo le consuetudini, normalmente si configura il servente HTTP in modo da non consentire la lettura del contenuto delle directory. In pratica, se si indica un indirizzo che rappresenta una directory, si ottiene invece un file predefinito, corrispondente di solito a '`index.html`' (o qualcosa di simile), contenuto nella directory richiesta; tuttavia, se questo è assente, non si ottiene alcunché.

Per poter usufruire di un servizio HTTP occorre un programma cliente adatto. In generale, tale programma cliente è in grado di accedere anche ad altri servizi, pertanto, in questo senso viene definito semplicemente «navigatore». Il programma di navigazione tipico dovrebbe consentire anche la visualizzazione di immagini e la fruizione di altri contenuti multimediali, ma un buon programma che utilizza soltanto un terminale a caratteri può funzionare in qualunque condizione, quindi, tale possibilità non deve essere scartata a priori.

Riquadro 40.1. *Uniform resource locator*, *Uniform resource identifier*

L'integrazione di diversi protocolli impone l'utilizzo di un sistema uniforme per indicare gli indirizzi, per poter conoscere subito in che modo si deve effettuare il collegamento. Per questo, quando si utilizza un navigatore, si devono usare indirizzi espressi in modo standard, precisamente secondo il formato URI, o *Uniform resource identifier*. Attualmente, è ancora in uso la vecchia definizione, URL, *Uniform resource locator*, la quale rappresenta un sottoinsieme di URI. Attraverso questa modalità, è possibile definire tutto quello che serve per raggiungere una risorsa: protocollo, nodo di rete (*host*), porta, percorso. Il formato generale di un URI è descritto nella sezione 54.1.

40.1 W3M

W3M[1] è un navigatore fatto per i terminali a caratteri, senza grafica, che però funziona correttamente con la codifica UTF-8 e ha una buona resa delle tabelle.

W3M si compone in pratica dell'eseguibile '**w3m**':

```
w3m [opzioni] risorsa_iniziale
```

Il file iniziale va indicato in forma di URI; eventualmente, se si tratta di file locali si può indicare il percorso senza URI.

Durante il funzionamento di W3M, la navigazione con la tastiera è abbastanza intuitiva e sono disponibili anche altri comandi molto interessanti. Si veda la tabella 40.2.

Tabella 40.2. Alcuni dei comandi di W3M.

Tastiera	Descrizione
[*H*]	Mostra il riepilogo dei comandi di navigazione.
[*pagina-su*], [*b*], [*Esc*][*v*]	Fa scorrere il testo all'indietro di una schermata.
[*pagina-giù*], [*spazio*], [*Ctrl v*]	Fa scorrere il testo in avanti di una schermata.
[*freccia-su*], [*Ctrl p*], [*k*]	Porta il cursore sulla riga precedente.
[*freccia-giù*], [*Ctrl n*], [*j*]	Porta il cursore sulla riga successiva.
[*Ctrl b*], [*h*]	Porta il cursore sul carattere precedente.
[*Ctrl f*], [*l*]	Porta il cursore sul carattere successivo.
[*J*]	Fa scorrere il testo in alto di riga.
[*K*]	Fa scorrere il testo in basso di riga.
[<]	Fa scorrere il testo verso sinistra.

Tastiera	Descrizione
[>]	Fa scorrere il testo verso destra.
[,]	Fa scorrere il testo verso sinistra di una sola colonna.
[.]	Fa scorrere il testo verso destra di una sola colonna.
[*Inizio*], [*g*]	Raggiunge l'inizio del file.
[*Fine*], [*G*]	Raggiunge la fine del file.
[*Ctrl u*], [*Esc*][*Tab*]	Raggiunge il riferimento ipertestuale precedente.
[*Tab*]	Raggiunge il riferimento ipertestuale successivo.
[*Invio*]	Accede al riferimento ipertestuale su cui si trova il cursore.
[*a*], [*Esc*][*Invio*]	Salva il riferimento ipertestuale in un file.
[*u*]	Mostra alla base dello schermo l'URI corrispondente al riferimento ipertestuale su cui si trova il cursore.
[*c*]	Mostra alla base dello schermo l'URI attuale.
[=]	Mostra le informazioni sull'URI attuale.
[*F*]	Cerca di elaborare una cornice (*frame*).
[*U*]	Permette di accedere a un URI da inserire manualmente.
[*F*]	Permette di accedere a un file da indicare manualmente.
[*R*]	Ricarica il documento.
[*S*]	Salva il documento su un file, come lo si vede sullo schermo.
[*Esc*][*s*]	Salva il documento su un file in forma originale.
[*s*]	Seleziona uno dei documenti visitati di recente.
[*o*]	Accede a una maschera di opzioni di funzionamento.

Tastiera	Descrizione
[*Ctrl s*], [*/*]	Richiede l'inserimento di una stringa di ricerca nella pagina attuale.
[*Ctrl r*], [*?*]	Come [*/*], ma ricerca all'indietro.
[*n*]	Continua la ricerca in avanti.
[*N*]	Continua la ricerca all'indietro.
[*q*]	Conclude il funzionamento chiedendo conferma.
[*Q*]	Conclude il funzionamento senza chiedere conferma.

W3M può essere avviato utilizzando diverse opzioni nella riga di comando. Tuttavia, di solito queste non si usano, potendo intervenire nel suo funzionamento attraverso il comando [*o*].

40.2 Servente HTTP: Mathopd

Mathopd[2] è un servente HTTP fatto per impegnare poche risorse, offrendo un insieme ragionevole di possibilità di configurazione.

Mathopd, da solo, non è in grado di mostrare il contenuto delle directory, in mancanza di un indice, inoltre produce un registro (log) che non è conforme agli standard, costituito di solito dal formato CLF (*Common log format*) o da quello combinato (sezione 40.7), ma è possibile rimediare a queste carenze con degli script o dei piccoli programmi di contorno.

Mathopd si compone del programma eseguibile '**`mathopd`**' che richiede un file di configurazione, corrispondente normalmente al file '`/etc/mathopd.conf`'. Il programma è fatto per funzionare da

solo, fuori dal controllo del supervisore dei servizi di rete, senza bisogno di avviare altre copie di se stesso.

40.2.1 Utilizzo generale

«

Mathopd è un servente HTTP molto «particolare», a cominciare dalla sintassi per l'avvio del programma '**mathopd**':

```
mathopd [opzioni] -f file_di_configurazione
```

Come si può osservare dal modello sintattico proposto, risulta obbligatorio indicare il file di configurazione con l'opzione '**-f**', perché in mancanza di questa informazione, il programma si aspetta di ricevere la configurazione dallo standard input.

Attraverso le altre opzioni che si trovano descritte nella pagina di manuale *mathopd(8)* è possibile controllare il funzionamento del servente per obbligarlo a funzionare in primo piano o a fornire informazioni diagnostiche. Attraverso una serie di segnali, è possibile attivare e disattivare delle funzionalità diagnostiche o intervenire sugli accessi in corso. In particolare, se il programma servente riceve il segnale SIGHUP rilegge la configurazione, mentre con SIGTERM o SIGINT termina di funzionare. A questo proposito, in un sistema GNU/Linux il servizio potrebbe essere controllato con uno script simile all'esempio seguente:

```
#! /bin/sh

case "$1" in
    start)
        echo "Avvio del servizio HTTP."
        /usr/sbin/mathopd -f /etc/mathopd.conf
        ;;
```

```
    stop)
        echo "Arresto del servizio HTTP."
        killall -s SIGTERM mathopd
        ;;
    reload)
        echo "Rilettura della configurazione del servizio"
        echo "HTTP."
        killall -s SIGHUP mathopd
        ;;
    *)
        echo "Utilizzo:"
        echo "/etc/init.d/mathopd {start|stop|reload}"
        exit 1
esac

exit 0
```

Durante il suo funzionamento, Mathopd ha la necessità di scrivere su tre file, che in condizioni normali coincidono con l'elenco seguente; tuttavia, si può modificare la collocazione e il nome di questi file intervenendo nella configurazione:

'`/var/run/mathopd.pid`'	contiene il numero del processo elaborativo (PID);
'`/var/mathopd/access.log`'	registro degli accessi;
'`/var/mathopd/error.log`'	registro degli errori.

A questo punto, sapendo che Mathopd annota il numero del processo elaborativo nel file '`/var/run/mathopd.pid`', o in qualunque altro file specificato nella configurazione, si può migliorare lo script di controllo del servizio in questo modo, rendendolo adatto a

un sistema GNU qualsiasi:

```
#! /bin/sh

case "$1" in
    start)
        echo "Avvio del servizio HTTP."
        /usr/sbin/mathopd -f /etc/mathopd.conf
        ;;
    stop)
        echo "Arresto del servizio HTTP."
        kill -s SIGTERM `cat /var/run/mathopd.pid`
        ;;
    reload)
        echo "Rilettura della configurazione del servizio"
        echo "HTTP."
        kill -s SIGHUP `cat /var/run/mathopd.pid`
        ;;
    *)
        echo "Utilizzo:"
        echo "/etc/init.d/mathopd {start|stop|reload}"
        exit 1
esac

exit 0
```

40.2.2 Configurazione

Come già spiegato, non esiste una posizione prestabilita del file di configurazione, cosa che deve essere specificata obbligatoriamente attraverso la riga di comando. Tuttavia, una posizione abbastanza logica per collocare questa configurazione è costituita dal file ‘`/etc/mathopd.conf`’, a cui si fa riferimento in generale nel capitolo; inoltre la pagina di manuale che descrive la sintassi di questo file

dovrebbe essere *mathopd.conf(5)*.

Il file di configurazione è un file di testo in cui le righe bianche o vuote vengono ignorate, così come viene ignorato il testo di una riga che appare dopo il simbolo '**#**'. Le direttive possono essere «semplici», a indicare ognuna l'attribuzione di un valore a un certo parametro di funzionamento, oppure possono essere dei blocchi di direttive. Un blocco, a sua volta, può contenere sia direttive semplici, sia blocchi ulteriori:

```
nome   valore_attribuito
```

```
nome  {
        direttiva
        ...
}
```

Come si può intendere, il primo modello si riferisce a una direttiva semplice, mentre il secondo mostra la dichiarazione di un blocco. Naturalmente, le parentesi graffe del secondo modello sintattico servono a delimitare l'insieme di direttive contenute nel blocco, pertanto sono da intendersi in senso letterale.

Ci sono direttive semplici che possono stare da sole senza essere inserite in un blocco particolare, mentre nella maggior parte dei casi, queste direttive semplici hanno valore solo nel contesto di un blocco specifico. Tutto questo è comunque abbastanza intuitivo, pertanto si intende mostrare qui la configurazione solo attraverso degli esempi; per approfondire la questione si deve leggere la pagina di manuale *mathopd.conf(5)*.

```
1  Umask 026
2
3  Tuning {
4          NumConnections 64
5          BufSize 12288
6          InputBufSize 2048
7          ScriptBufSize 4096
8          NumHeaders 100
9          Timeout 240
10         ScriptTimeout 120
11 }
12
13 User www-data
14 StayRoot Off
15
16 PIDFile /var/run/mathopd.pid
17 Log /var/log/mathopd/access.log
18 ErrorLog /var/log/mathopd/error.log
19
20 Control {
21         ScriptUser nobody
22         ChildLog /var/log/mathopd/child.log
23         Types {
24                 text/html { html htm }
25                 text/plain { txt }
26                 image/gif { gif }
27                 image/jpeg { jpg }
28                 image/png { png }
29                 text/css { css }
30                 audio/midi { mid midi kar }
31                 application/octet-stream { * }
32         }
33         External {
34                 /usr/bin/php { php }
```

```
35          }
36          IndexNames { index.html index.htm }
37  }
38
39  Server {
40          Port 80
41          Address 0.0.0.0
42          Virtual {
43                  AnyHost
44                  Control {
45                          Alias /
46                          Location /var/www
47                          Access {
48                                  Allow 0/0
49                          }
50                  }
51                  Control {
52                          Alias /cgi-bin
53                          Location /usr/lib/cgi-bin
54                          Specials {
55                                  CGI { * }
56                          }
57                          Access {
58                                  Allow 0/0
59                          }
60                  }
61                  Control {
62                          Alias /~
63                          Location public_html
64                          UserDirectory On
65                          RunScriptsAsOwner On
66                          Access {
67                                  Allow 0/0
68                          }
```

```
69                      }
70                      Control {
71                              Alias /~root
72                              Location /nosuchdirectory
73                              Access {
74                                      Deny 0/0
75                                      Allow 127.0.0.1/32
76                              }
77                      }
78                      Control {
79                              Alias /doc
80                              Location /usr/share/doc
81                              Access {
82                                      Deny 0/0
83                                      Allow 127.0.0.1/32
84                              }
85                      }
86                      Control {
87                              Alias /dwww
88                              Location /var/lib/dwww/html
89                              Access {
90                                      Deny 0/0
91                                      Allow 127.0.0.1/32
92                              }
93                      }
94              }
95      }
```

L'esempio appena mostrato riguarda una situazione abbastanza comune, dove si gestisce un solo dominio virtuale e il materiale pubblicato è generalmente disponibile a tutti. Per maggiore comodità, l'esempio viene sezionato durante la sua descrizione.

```
1       Umask 026
```

Questa direttiva iniziale, che non è racchiusa in alcun gruppo, dichiara la maschera dei permessi che si vuole sia usata per i file che Mathopd va a creare. In questo caso, viene tolto il permesso di scrittura al gruppo (2_8) e vengono tolti i permessi di lettura e scrittura agli utenti che non sono né il proprietario del file, né gli utenti del gruppo a cui questo è associato (6_8). In pratica, sapendo che non può entrare in gioco il permesso di esecuzione, il proprietario può leggere e modificare i file, mentre il gruppo può solo leggere.

```
3   Tuning {
4           NumConnections 64
5           BufSize 12288
6           InputBufSize 2048
7           ScriptBufSize 4096
8           NumHeaders 100
9           Timeout 240
10          ScriptTimeout 120
11  }
```

Il raggruppamento denominato '**Tuning**' consente di inserire alcune direttive che regolano il funzionamento generale. Il significato di queste può risultare abbastanza intuitivo; in particolare viene definito il numero massimo di connessioni simultanee (in questo caso sono 64) e la scadenza, sia per le connessioni, sia per l'esecuzione di un programma CGI (nell'esempio, le connessioni scadono dopo 240 s, mentre i programmi CGI devono concludersi entro 120 s).

Sulla base dei valori assegnati a queste direttive, è possibile calcolare la quantità di memoria utilizzata da Mathopd:

$$NumConnections \cdot BufSize + InputBufSize + 2 \cdot ScriptBufSize$$

```
13  User www-data
```

Quando Mathopd viene avviato con i privilegi dell'utente '**root**', si deve utilizzare questa direttiva per fare in modo che, subito dopo l'avvio, il programma servente passi ai privilegi dell'utente indicato. In questo modo, tra le altre cose, i file che Mathopd utilizza devono essere accessibili a tale utente. Questo problema vale sia per i documenti da pubblicare, sia per i programmi da eseguire, sia per i file delle registrazioni. Il gruppo non viene specificato e questo dipende dal tipo di adattamento particolare di Mathopd (in un sistema GNU dovrebbe trattarsi del gruppo abbinato naturalmente all'utente indicato).

```
14  StayRoot Off
```

Questa direttiva, se attiva, fa sì che alcune funzioni di Mathopd vengano eseguite con i privilegi dell'utente '**root**', nonostante sia usata la direttiva '**User**'. In certi casi, ciò può essere utile, ma in generale è meglio evitare questo.

```
16  PIDFile /var/run/mathopd.pid
17  Log /var/log/mathopd/access.log
18  ErrorLog /var/log/mathopd/error.log
```

Queste direttive permettono di stabilire la collocazione dei file usati per annotare il numero PID del programma servente e per i file delle registrazioni.

```
20  Control {
21          ScriptUser nobody
22          ChildLog /var/log/mathopd/child.log
```

Il gruppo '**Control**' serve a raggruppare delle direttive che controllano il comportamento del servente. Quando il gruppo si trova in un contesto generale (al di fuori di qualunque altro blocco), le direttive valgono per ogni situazione, salva la possibilità di ridefinire i

parametri in contesti più specifici.

All'inizio del gruppo '**Control**' si vedono due direttive; la prima dichiara con quali privilegi debbano essere eseguiti i programmi CGI, ma per funzionare è necessario che la direttiva '**StayRoot**' sia attiva; pertanto, in questo caso la richiesta di eseguire i programmi CGI con i privilegi dell'utente '**nobody**' non può essere soddisfatta. La seconda direttiva che si vede dichiara un file nel quale annotare quanto emesso attraverso lo standard error dai programmi CGI. In mancanza di questa direttiva, tali messaggi vengono perduti (la parola *child* fa riferimento al fatto che i programmi CGI sono processi elaborativi discendenti da quello del servente).

```
23          Types {
24                  text/html { html htm }
25                  text/plain { txt }
26                  image/gif { gif }
27                  image/jpeg { jpg }
28                  image/png { png }
29                  text/css { css }
30                  audio/midi { mid midi kar }
31                  application/octet-stream { * }
32          }
```

Il gruppo '**Types**' è necessario per dichiarare i tipi di file in base all'estensione. Come si può vedere, i file HTML vengono riconosciuti in base all'estensione '**html**' o anche solo '**htm**'. L'ultima direttiva di questo gruppo deve indicare un tipo adatto a descrivere i file che hanno estensioni differenti da quelle previste espressamente (l'asterisco serve a indicare qualunque estensione). Purtroppo, questo è un limite importante di Mathopd, non essendo in grado di individuare i file di testo senza estensione, a meno di usare tale dichiarazione per

ultima. Per la precisione, l'estensione indicata non implica automaticamente la presenza di un punto, pertanto, può essere più corretto aggiungere questo punto nell'estensione stessa. A titolo di esempio, l'elenco dei tipi potrebbe essere esteso come nell'estratto seguente:

```
Types {
    application/ogg                          { .ogg }
    application/pdf                          { .pdf }
    application/postscript                   { .ps .ai .eps }
    application/rtf                          { .rtf }
    application/xhtml+xml                    { .xht .xhtml }
    application/zip                          { .zip }
    application/x-cpio                       { .cpio }
    application/x-debian-package             { .deb }
    application/x-dvi                        { .dvi }
    application/x-gtar                       { .gtar .tgz .taz }
    application/x-redhat-package-manager     { .rpm }
    application/x-tar                        { .tar }
    audio/midi                               { .mid .midi .kar }
    audio/mpeg                               { .mpga .mpega
                                               .mp2 .mp3 .m4a }
    audio/x-mpegurl                          { .m3u }
    audio/x-wav                              { .wav }
    image/gif                                { .gif }
    image/jpeg                               { .jpeg .jpg .jpe }
    image/pcx                                { .pcx }
    image/png                                { .png }
    image/tiff                               { .tiff .tif }
    text/css                                 { .css }
    text/html                                { .htm .html .shtml }
    text/plain                               { .asc .txt .text
                                               .diff .pot
                                               readme README
                                               LEGGIMI
                                               COPYRIGHT
                                               COPYING }
    text/rtf                                 { .rtf }
```

```
    text/xml                                { .xml .xsl }
    video/fli                               { .fli }
    video/mpeg                              { .mpeg .mpg .mpe .mp4 }
    video/quicktime                         { .qt .mov }
    video/x-ms-asf                          { .asf .asx }
    video/x-msvideo                         { .avi }

    application/octet-stream                { * }
}
```

Evidentemente, date le caratteristiche di Mathopd, conviene estendere questo elenco solo quando si presenta la necessità, in base ai contenuti dei documenti pubblicati.

```
33          External {
34                  /usr/bin/php { php }
35          }
```

Il gruppo '**`External`**' serve a delimitare delle direttive che dichiarano l'uso di un programma interprete per eseguire i file con le estensioni indicate. In questo caso, quando si incontra un file con estensione '`php`', questo viene eseguito attraverso il programma '`/usr/bin/php`'. Come già per le direttive del gruppo '**`Types`**', può essere più conveniente aggiungere il punto che precede l'estensione, come nell'esempio seguente dove però vengono aggiunte altre estensioni equivalenti:

```
        External {
                /usr/bin/php    { .php .phtml .pht }
        }
```

Si osservi che per quanto riguarda gli script che hanno i permessi per essere eseguibili, si attivano attraverso un'altra direttiva nel

gruppo '**Specials**', come nell'esempio successivo, che suppone si inserisca all'interno del gruppo '**Control**' principale:

```
        Specials {
                CGI             { .cgi .sh .pl }
        }
```

Come si può intuire, in questo esempio si intenderebbe dichiarare come programmi esterni i file che terminano per '`.cgi`', '`.sh`' e '**`.pl`**'. Nell'esempio complessivo questo caso è stato escluso, per dichiarare piuttosto l'uso del gruppo '**Special**' nell'ambito di un percorso specifico.

```
36          IndexNames { index.html index.htm }
```

Questa direttiva dichiara quali file usare come indici delle directory.

```
39  Server {
40          Port 80
41          Address 0.0.0.0
```

Il gruppo '**Server**' contiene direttive riferite a un servente HTTP in ascolto in una certa porta, per tutti o solo per un certo indirizzo IP. Nell'esempio si attiva un servente in ascolto della porta 80, il quale accetta connessioni da qualunque indirizzo IPv4.

```
42          Virtual {
43                  AnyHost
```

Il gruppo '**Virtual**' serve a delimitare un insieme di direttive relativo a un certo dominio virtuale. In questo caso, con la direttiva '**AnyHost**' si specifica che il gruppo riguarda qualunque dominio che non sia stato individuato in modo più dettagliato.

```
44                      Control {
45                              Alias /
46                              Location /var/www
47                              Access {
48                                      Allow 0/0
49                              }
50                      }
```

All'interno dei gruppi '**`Virtual`**' si indicano dei gruppi '**`Control`**' per individuare dei percorsi, a cui associare dei comportamenti. In questo caso, si dichiara il percorso iniziale del dominio, che corrisponde nel file system alla directory '`/var/www/`'. Come si può intuire, nel gruppo '**`Access`**' viene concesso espressamente l'accesso da qualunque indirizzo.

```
51                      Control {
52                              Alias /cgi-bin
53                              Location /usr/lib/cgi-bin
54                              Specials {
55                                      CGI { * }
56                              }
57                              Access {
58                                      Allow 0/0
59                              }
60                      }
```

Il gruppo '**`Control`**' successivo nell'esempio iniziale, ha lo scopo di associare il percorso ***dominio*/cgi-bin/** alla directory locale '`/usr/lib/cgi-bin/`', specificando che ogni file contenuto al suo interno è da intendere un programma CGI. Anche in questo caso viene concesso l'accesso a chiunque.

```
61                      Control {
62                              Alias /~
63                              Location public_html
64                              UserDirectory On
65                              RunScriptsAsOwner On
66                              Access {
67                                      Allow 0/0
68                              }
69                      }
70                      Control {
71                              Alias /~root
72                              Location /nosuchdirectory
73                              Access {
74                                      Deny 0/0
75                                      Allow 127.0.0.1/32
76                              }
77                      }
```

Qui si dichiara l'accessibilità alla directory personale di ogni utente. Come si fa normalmente, gli accessi riguardano precisamente la directory '`~/public_html/`' e ciò che questa contiene. Teoricamente, in base alla direttiva '**`RunScriptsAsOwner`**', che risulta attiva, i programmi CGI contenuti all'interno della gerarchia degli utenti, dovrebbero essere eseguiti con i privilegi degli utenti stessi. In pratica, dal momento che in precedenza il parametro associato alla direttiva '**`StayRoot`**' è stato disattivato, l'attivazione di '**`RunScriptsAsOwner`**' diventa priva di significato.

Per evitare di trattare nello stesso modo anche l'utente '**`root`**', viene dichiarato un gruppo apposito, dove il percorso *http://**nodo**/~root/* viene associato deliberatamente a una directory inesistente, per garantire che non vi si possa accedere. Sotto, con il gruppo '**`Access`**'

viene escluso ogni accesso, salvo all'elaboratore locale, ma per il motivo appena descritto risulta ugualmente inaccessibile.

```
78                  Control {
79                          Alias /doc
80                          Location /usr/share/doc
81                          Access {
82                                  Deny 0/0
83                                  Allow 127.0.0.1/32
84                          }
85                  }
86                  Control {
87                          Alias /dwww
88                          Location /var/lib/dwww/html
89                          Access {
90                                  Deny 0/0
91                                  Allow 127.0.0.1/32
92                          }
93                  }
```

Infine, vengono previsti altri percorsi a directory contenenti della documentazione. A questi percorsi viene impedito l'accesso a tutti, escluso l'elaboratore locale.

Per dichiarare dei domini virtuali, si potrebbe continuare con altri gruppi '**`Virtual`**' che iniziano con una o più direttive '**`Host`**', come nell'esempio seguente:

```
        Virtual {
                Host brot.dg
                Host www.brot.dg
                Control {
                        Alias /
                        Location /home/MIRROR/brot
                }
```

```
                Control {
                        Alias /cgi-bin
                        Location /nosuchdirectory
                        Access {
                                Deny 0/0
                        }
                }
                Control {
                        Alias /~
                        Location /nosuchdirectory
                        Access {
                                Deny 0/0
                        }
                }
                Control {
                        Alias /~root
                        Location /nosuchdirectory
                        Access {
                                Deny 0/0
                        }
                }
        }
```

40.2.3 Indici delle directory

Purtroppo, Mathopd non consente di visualizzare il contenuto di un percorso nel quale non è stato previsto un indice. Tuttavia, se si dispone di un programma CGI che genera l'indice, è possibile collocare tale programma in ogni directory priva di un altro indice e abilitarne l'uso nella configurazione:

```
Control {
        ScriptUser nobody
```

```
        ChildLog /var/log/mathopd/child.log
        Types {
                …
        }
        Specials {
                CGI { dir_cgi }
        }
        IndexNames { index.html index.htm dir_cgi }
}
```

Come si vede, nel gruppo '**Control**' più esterno si può inserire un gruppo '**Specials**' allo scopo di dichiarare «l'estensione» '`dir_php`' come programma CGI, mettendo lo stesso nome nell'elenco dei file indice.

In pratica, non si tratta di un'estensione, ma del nome del file completo: se al posto del file '`index.html`', o '`index.htm`', c'è il programma '**dir_cgi**', questo viene eseguito.

Il nome '**dir_cgi**' non è casuale, in quanto si tratta di un esempio diffuso dallo stesso autore di Mathopd (una copia di questo programma in forma sorgente, dovrebbe essere disponibile anche qui: *allegati/dir_cgi.c*).

Un risultato simile si può ottenere con il programma *allegati/index-html.cgi* che è scritto in Perl.

40.2.4 Registro degli accessi

Il formato usato da Mathopd per annotare gli accessi nel file '`/var/log/mathopd/access.log`', o comunque nel file equivalente stabilito in base alla configurazione, non è standard. Nella configurazio-

ne si può intervenire con una serie di direttive racchiuse nel gruppo '**LogFormat**':

```
LogFormat {
        Ctime
        RemoteUser
        RemoteAddress
        RemotePort
        ServerName
        Method
        URI
        Status
        ContentLength
        Referer
        UserAgent
        BytesRead
        BytesWritten
}
```

Quello che si ottiene è un file di testo, contenente delle righe, una per ogni richiesta giunta al servente, in cui le varie informazioni sono separate da un carattere di tabulazione orizzontale (<*HT*>). L'esempio mostrato sopra nell'uso del gruppo '**LogFormat**', rappresenta la sequenza dei campi predefiniti; tuttavia, anche cambiando la disposizione di questi campi, non si può ottenere il formato CLF (*Common log format*) e tanto meno quello combinato (sezione 40.7). Per disporre di un formato standard, è necessari rielaborare il file con un programma realizzato appositamente, pertanto è perfettamente inutile modificare la disposizione dei campi nella configurazione di Mathopd.

Nei punti di distribuzione di Mathopd potrebbero essere disponibili due script alternativi, che in qualche modo dovrebbero generare un

formato combinato da un file di registrazione degli accessi predefinito. Il primo di questi è uno script AWK (una copia di questo script dovrebbe essere disponibile anche qui: *allegati/mattoclf.awk*):

```
#!/usr/bin/awk -f
BEGIN {
        FS="\t"
}
NF >= 11 && $5 ~ "^[-._[:alnum:]]+$" {
        split($1, date, " ")
        printf "%s - - [%02d/%s/%s:%s +0000] \"%s %s HTTP/1.0\" %d %d \"%s\" \"%s\"\n",
            $3, date[3], date[2], date[5], date[4], $6, $7, $8, $9, $10, $11 > $5
}
```

Questo script attende dallo standard input il contenuto del registro degli accessi e genera tanti file quanti sono i domini virtuali. Ognuno di questi file, ha il nome del dominio virtuale relativo.

Questo programma sarebbe perfetto, se non fosse che, quando manca l'informazione del dominio virtuale (pertanto appare in quella posizione un trattino), si blocca, perché non può creare il file '–'.

Un altro script, questa volta in Perl, fa un lavoro simile, ma senza distinguere tra i domini virtuali (una copia di questo script dovrebbe essere disponibile anche qui: *allegati/mattoclf.pl*):

```
#!/usr/bin/perl

while(<STDIN>) {
        my(@x) = split(/\t/);
        my $date = shift @x;
        my @date = split(/\s+/, $date);
        if ($x[8] eq '-') {
                $x[8] = '';
        }
        printf "%s - - [%02d/%s/%s:%s +0000] \"%s %s HTTP/1.0\" %d %d \"%s\" \"%s\"\n",
                $x[1], $date[2], $date[1], $date[4], $date[3], $x[4], $x[5], $x[6],
                $x[7], $x[8], $x[9];
}
```

Data la mancanza di un programma soddisfacente nella distribuzione di Mathopd, viene proposto qui un programma Perl differente, più completo, che genera un risultato equivalente a quello del programma AWK già apparso sopra, ma senza incepparsi quando manca il nome del dominio virtuale: *allegati/mathopd_to_clf* .

40.3 Protocollo HTTP

«

Il funzionamento del protocollo HTTP è molto semplice. L'utilizzo di un servizio HTTP si compone di una serie di transazioni, ognuna delle quali si articola in queste fasi:

1. apertura della connessione;
2. invio da parte del cliente di una richiesta;
3. risposta da parte del servente;
4. chiusura della connessione.

In questo modo, il programma servente non deve tenere traccia delle transazioni che iniziano e finiscono ogni volta che un utente compie un'azione attraverso il suo programma cliente.

La richiesta inviata dal programma cliente deve contenere il metodo (i più comuni sono '**GET**' e '**POST**'), l'indicazione della risorsa cui si vuole accedere, la versione del protocollo ed eventualmente l'indicazione dei tipi di dati che possono essere gestiti dal programma cliente (si parla in questi casi di tipi MIME). Naturalmente sono possibili richieste più ricche di informazioni.

Tabella 40.31. Alcuni metodi di comunicazione per le richieste di un programma cliente.

Nome	Descrizione
`GET`	Recupera l'informazione identificata dall'URI specificato.
`HEAD`	Recupera le informazioni sul documento, senza ottenere il documento in allegato.
`PUT`	Richiede che l'informazione sia memorizzata nell'URI specificato.
`POST`	Fornisce al servente HTTP dei dati aggiuntivi in riferimento all'URI specificato.
`DELETE`	Richiede di eliminare la risorsa specificata.
`LINK`	Stabilisce una collegamento con la risorsa indicata.
`UNLINK`	Elimina un collegamento tra risorse.

La risposta del servente HTTP è costituita da un'intestazione che, tra le altre cose, specifica il modo in cui l'informazione allegata deve essere interpretata. È importante comprendere subito che l'intestazione viene staccata dall'inizio dell'informazione allegata attraverso un riga vuota, composta dalla sequenza *<CR><LF>*.

40.3.1 Analisi di una connessione HTTP

«

Per comprendere in pratica il funzionamento di una connessione HTTP, si può utilizzare il programma '**telnet**' al posto di un navigatore normale. Si suppone di poter accedere al nodo *www.brot.dg* nel quale è stato installato un servente HTTP con successo. Dal servente viene prelevato il file '`index.html`' che si trova all'interno della directory principale del servizio, ovvero da *document root*.

```
$ telnet www.brot.dg http[Invio]
```

Il programma '**telnet**' risponde e si mette in attesa di ricevere il messaggio da inviare al servente:

```
Trying 192.168.1.1...
Connected to www.brot.dg.
Escape character is '^]'.
```

Si deve iniziare a scrivere, cominciando con una riga contenente il metodo, la risorsa e la versione del protocollo, continuando con una riga contenente le possibilità di visualizzazione del cliente (i tipi MIME).

```
GET /index.html HTTP/1.0[Invio]

Accept: text/html[Invio]

[Invio]
```

Appena si invia una riga vuota, il servente intende che la richiesta è terminata e risponde:

```
HTTP/1.1 200 OK
Date: Tue, 27 Jan 1998 17:44:46 GMT
Server: Apache/1.2.4
```

```
Last-Modified: Tue, 30 Dec 1997 21:07:24 GMT
ETag: "6b003-792-34a9628c"
Content-Length: 1938
Accept-Ranges: bytes
Connection: close
Content-Type: text/html

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
<HTML>
 <HEAD>
  <TITLE>Test Page for Linux's Apache Installation</TITLE>
 </HEAD>
 <BODY
  BGCOLOR="#FFFFFF"
  TEXT="#000000"
  LINK="#0000FF"
  VLINK="#000080"
  ALINK="#FF0000"
 >
  <H1 ALIGN="CENTER">It Worked!</H1>
  <P>
  If you can see this, it means that the installation of the
  <A
   HREF="http://www.apache.org/"
  >Apache</A>
  software on this Linux system was successful. You may now
  add content to this directory and replace this page.
  </P>
...
...
 </BODY>
</HTML>
Connection closed by foreign host.
```

Come già accennato, il messaggio restituito dal servente è

composto da un'intestazione in cui l'informazione più importante è il tipo di messaggio allegato, cioè in questo caso '**`Content-Type: text/html`**', seguita da una riga vuota e quindi dall'oggetto richiesto, cioè il file '`index.html`'.

Al termine della ricezione dell'oggetto richiesto, la connessione ha termine. Lo si può osservare dal messaggio dato da '**`telnet`**': '**`Connection closed by foreign host`**'.

Il lavoro di un programma cliente è tutto qui: inviare richieste al servente HTTP, ricevere le risposte e gestire i dati, possibilmente visualizzandoli o mettendo comunque l'utente in grado di fruirne.

40.3.2 Tipi MIME

«

MIME è una codifica standard per definire il trasferimento di documenti multimediali attraverso la rete. L'acronimo sta per *Multipurpose Internet mail extentions* e la sua origine è appunto legata ai trasferimenti di dati allegati ai messaggi di posta, come il nome lascia intendere.

Il protocollo HTTP utilizza lo stesso standard e con questo il programma servente informa il programma cliente del tipo di oggetto che gli viene inviato. Nello stesso modo, il programma cliente, all'atto della richiesta di una risorsa, informa il servente dei tipi MIME che è in grado di gestire.

Il servente HTTP, per poter comunicare il tipo MIME al cliente, deve avere un modo per riconoscere la natura degli oggetti che costituiscono le risorse accessibili. Questo modo è dato solitamente dall'estensione, per cui, la stessa scelta dell'estensione per i file accessibili attraverso il protocollo HTTP è praticamente obbligatoria, ovvero, dipende dalla configurazione dei tipi MIME.

Tabella 40.34. Alcuni tipi MIME con le possibili estensioni.

Tipo MIME	Estensioni	Descrizione
application/postscript	ps eps	PostScript.
application/rtf	rtf	Rich Text Format.
application/x-tex	tex	Documento TeX/LaTeX.
audio/basic	au snd	File audio.
audio/x-wav	wav	File audio.
image/gif	gif	Immagine GIF.
image/jpeg	jpeg jpg	Immagine JPEG.
image/tiff	tiff tif	Immagine TIFF.
image/x-xwindowdump	xwd	Immagine X Window Dump.
text/html	html htm	Testo composto in HTML.
text/plain	txt	Testo puro.
video/mpeg	mpeg mpg mpe	Animazione MPEG.
video/quicktime	qt mov	Animazione Quicktime.

40.3.3 Campi di richiesta

Come si è visto dagli esempi mostrati precedentemente, la richiesta fatta dal programma cliente è composta da una prima riga in cui si dichiara il tipo, la risorsa desiderata e la versione del protocollo.

```
GET /index.html HTTP/1.0
```

Di seguito vengono indicati una serie di campi, più o meno facoltativi. Questi campi sono costituiti da un nome seguito da due punti (‘`:`’), da uno spazio e dall’informazione che gli si vuole abbinare.

Campo «Accept»

Una o più righe contenenti un campo ‘**`Accept`**’ possono essere

incluse per indicare i tipi MIME che il cliente è in grado di gestire (cioè di ricevere). Se non viene indicato alcun campo '**Accept**', si intende che siano accettati almeno i tipi '**text/plain**' e '**text/html**'.

I tipi MIME sono organizzati attraverso due parole chiave separate da una barra obliqua. In pratica si distingue un tipo e un sottotipo MIME. È possibile indicare un gruppo di tipi MIME mettendo un asterisco al posto di una o di entrambe le parole chiave, in modo da selezionare tutto il gruppo relativo. Per esempio,

```
Accept: */*
```

rappresenta tutti i tipi MIME;

```
Accept: text/*
```

rappresenta tutti i sottotipi MIME che appartengono al tipo '**text**'; mentre

```
Accept: audio/basic
```

rappresenta un tipo e un sottotipo MIME particolare.

Campo «User-Agent»

Il campo '**User-Agent**' permette di informare il servente sul nome e sulla versione dell'applicativo particolare che svolge la funzione di cliente. Per convenzione, il nome di questo è seguito da una barra obliqua e dal numero della versione. Tutto quello che dovesse seguire sono solo informazioni addizionali per le quali non è stabilita una forma precisa. Per esempio, nel caso di Mozilla, si potrebbe avere un'indicazione del tipo seguente:

```
User-Agent: Mozilla/4.04 [en] (X11; I; Linux 2.0.32 i586)
```

40.3.4 Campi di risposta

La risposta del servente HTTP a una richiesta del programma cliente si compone di un'intestazione seguita eventualmente da un allegato, il quale costituisce la risorsa a cui il cliente voleva accedere. L'intestazione è separata dall'allegato da una riga vuota.

La prima riga è costituita dal codice di stato della risposta. Nella migliore delle ipotesi dovrebbe presentarsi come nell'esempio seguente:

```
HTTP/1.0 200 OK
```

Tabella 40.41. Alcuni codici di stato utilizzati più frequentemente.

Codice	Descrizione	Codice	Descrizione
200	OK	201	Creato.
202	Accettato.	204	Nessun contenuto.
300	Scelte multiple.	301	Spostato in modo permanente.
302	Spostato temporaneamente.	304	Non modificato.
400	Richiesta errata.	401	Non autorizzato.
403	Proibito.	404	Non trovato.
500	Errore interno del servente HTTP.	501	Servizio non realizzato (non disponibile).
502	Gateway errato.	503	Servizio non disponibile.

Il resto dell'intestazione è composto da campi, simili a quelli utilizzati per le richieste dei programmi clienti.

Campo «Allow»

Il campo '**`Allow`**' viene utilizzato dal programma servente per informare il programma cliente dei metodi che possono essere utilizzati. Viene restituita tale informazione quando il cliente tenta di utilizzare un metodo di richiesta che il servente non è in grado di gestire. Segue un esempio.

```
Allow: GET, HEAD, POST
```

Campo «Content-Length»

Il campo '**`Content-Length`**' indica al programma cliente la dimensione (in byte) dell'allegato. Se viene utilizzato il metodo '**`HEAD`**', con cui non viene restituito alcun allegato, permette di conoscere in anticipo la dimensione della risorsa.

```
Content-Length: 1938
```

Campo «Content-Type»

Il campo '**`Content-Type`**' indica al programma cliente il tipo MIME a cui appartiene la risorsa (allegata o meno). Segue l'esempio più comune.

```
Content-Type: text/html
```

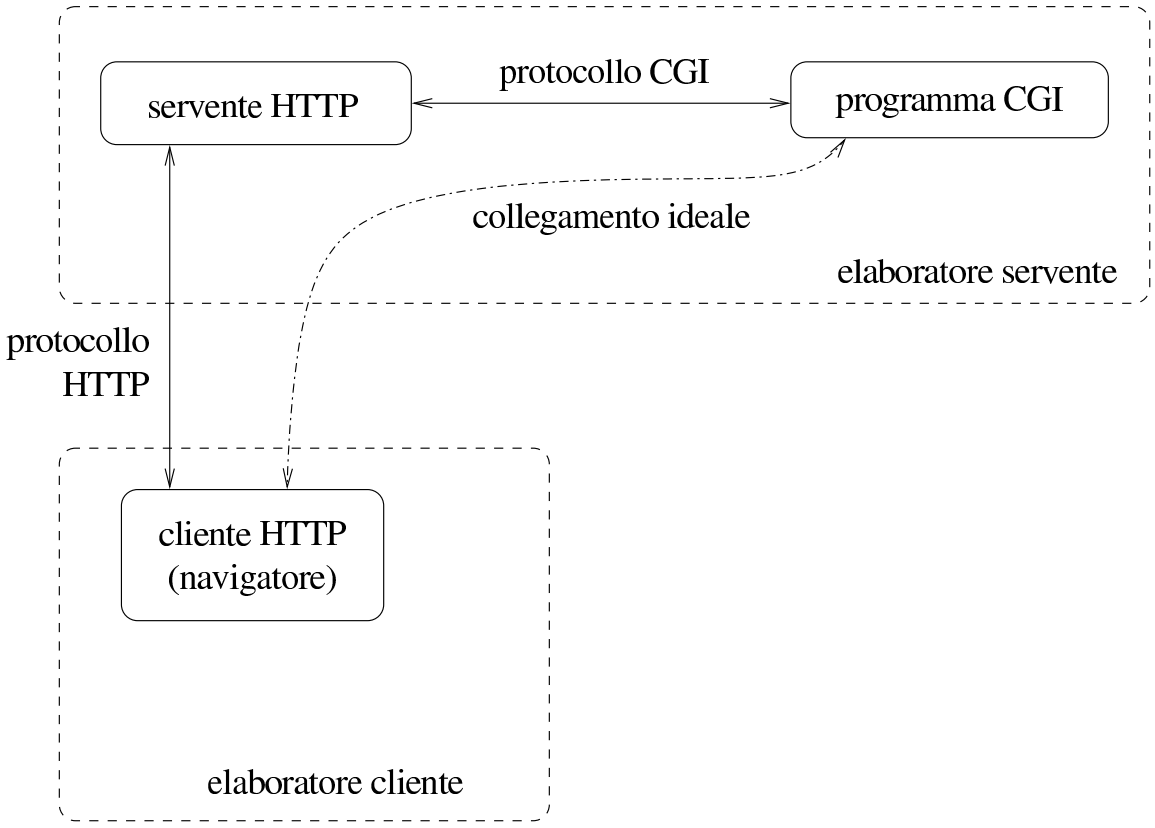
40.4 HTTP e CGI

«

HTTP (*Hypertext transfer protocol*) è un protocollo cliente-servente progettato per gestire documenti ipertestuali e per permettere l'interazione con programmi, detti ***gateway***, attraverso le specifiche CGI (*Common gateway interface*).

L'interfaccia CGI permette quindi di realizzare programmi che interagiscono con gli utenti attraverso il protocollo HTTP. La figura 40.45 illustra il meccanismo.

Figura 40.45. Schema del collegamento fisico e ideale tra le varie parti di una connessione HTTP-CGI.



I programmi *gateway*, detti anche *cgi-bin* o più semplicemente CGI, possono essere realizzati con qualunque linguaggio, purché siano in grado di interagire attraverso le specifiche del protocollo CGI.

40.4.1 URI e query

Vale la pena di richiamare brevemente alcuni concetti riferiti agli URI per ciò che riguarda in particolare la gestione interattiva che si vuole descrivere in questo capitolo (si veda eventualmente la sezione 54.1). Il formato di un URI potrebbe essere definito secondo lo schema seguente:

protocollo ***indirizzo_della_risorsa*** [***dati_aggiuntivi***]

Alcuni tipi di protocolli sono in grado di gestire dei dati aggiuntivi in coda all'indirizzo della risorsa. Nel caso del protocollo HTTP combinato con CGI, può trattarsi di ***richieste*** o di ***percorsi aggiuntivi***.

Quando un URI comprende anche una stringa di richiesta (*query*), questa viene distinta dall'indirizzo della risorsa attraverso un punto interrogativo.[3]

protocollo ***indirizzo_della_risorsa*** ? [***richiesta***]

L'utilizzo di una stringa di richiesta presume che la risorsa sia un programma in grado di utilizzare l'informazione contenuta in tale stringa. Segue un esempio banale di un URI contenente una richiesta:

http://www.brot.dg/cgi-bin/saluti.pl?buongiorno

Quando l'indirizzo della risorsa di un URI fa riferimento a un programma, questo può ricevere un'informazione aggiuntiva legata a un file o a una directory particolare. Si ottiene questo aggiungendo l'indicazione del percorso che identifica questo file o questa directory.

protocollo ***indirizzo_della_risorsa*** [***percorso_aggiuntivo***]

Segue un esempio banale di un URI, completo dell'indicazione di un percorso:

http://www.brot.dg/cgi-bin/elabora.pl/archivio.doc

Quando un simbolo di quelli non utilizzabili deve essere indicato

ugualmente da qualche parte dell'URI, facendogli perdere il significato speciale che questo potrebbe avere altrimenti, si può convertire utilizzando la notazione '**%*hh***'. La sigla ***hh*** rappresenta una coppia di cifre esadecimali. A questa regola fa eccezione lo spazio che viene codificato normalmente con il segno '**+**', ma non in tutte le occasioni.

Generalmente, per gli indirizzi URI normali non c'è la necessità di preoccuparsi di questo problema, quindi, l'utilizzo di simboli particolari riguarda prettamente la costruzione delle richieste, come viene mostrato meglio in seguito.

La tabella 40.46 mostra l'elenco di alcune corrispondenze tra simboli particolari e la codifica alternativa utilizzabile negli URI.

Tabella 40.46. Alcune corrispondenze tra simboli particolari e codifica alternativa utilizzabile negli URI.

Carattere	Codifica	Carattere	Codifica
%	%25	&	%26
+	%2B	/	%2F
=	%3D	~	%7E

40.4.2 Input dell'utente

Il tipo di comunicazione che avviene tra programma cliente e programma servente, descritta in precedenza, è nascosta all'utente, il quale agisce attraverso la richiesta e l'invio di documenti HTML. Si distinguono tre tipi di definizioni da inserire all'interno di documenti HTML che permettono all'utente di inserire dati (nel senso di input): elemento '**ISINDEX**' superato e destinato a essere eliminato dallo standard; attributo '**ISMAP**' delle immagini, ma anche questo superato; elementi '**FORM**'. Considerato che tutto quello che si potrebbe

fare con gli elementi '**ISINDEX**' e gli attributi '**ISMAP**', si può fare con gli elementi '**FORM**' e il loro contenuto, è meglio concentrarsi su questa ultima possibilità.

Gli elementi '**FORM**' consentono genericamente di realizzare dei ***formulari***, ovvero delle maschere o modelli per l'inserimento di dati. Le informazioni fornite dall'utente in questo modo vengono trasmesse nella forma di stringhe che rappresentano l'assegnamento di un valore a una variabile: '***nome=valore***'. I dati inseriti attraverso gli elementi '**FORM**' possono essere trasmessi con una richiesta '**GET**' oppure '**POST**', attraverso l'indicazione opportuna all'interno dello stesso documento HTML che contiene il formulario. La descrizione di questi elementi '**FORM**' viene fatta più avanti.

40.4.3 Primo approccio alla programmazione CGI

«

I programmi *gateway*, o CGI, vengono visti dai clienti come delle risorse normali. Alla chiamata, tali programmi restituiscono, attraverso il servente, un documento HTML. I programmi *gateway* generano del codice HTML e lo emettono attraverso lo standard output, il quale viene intercettato dal servente, il quale a sua volta lo completa inizialmente del codice di stato. In pratica, un programma del genere riceve input in qualche modo attraverso il servente, il quale a sua volta ha ricevuto una richiesta da un cliente, quindi restituisce un documento HTML preceduto da un'intestazione, ma senza la riga di stato.

Un programma CGI banale, potrebbe essere quello che restituisce semplicemente un messaggio composto in HTML, ogni volta che viene eseguito (una copia di questo file dovrebbe essere disponibile anche qui: *allegati/cgi-banale.sh*).

```
#!/bin/sh

echo "Content-type: text/html"
echo
echo "<HTML>"
echo "<HEAD>"
echo "<TITLE>Programma CGI banale</TITLE>"
echo "</HEAD>"
echo "<BODY>"
echo "<H1>Programma CGI banale</H1>"
echo "<P>"
echo "Ciao Mondo!"
echo "</P>"
echo "</BODY>"
echo "</HTML>"
```

Supponendo di avere chiamato questo programma '**`cgi-banale.sh`**' e di averlo reso eseguibile, supponendo inoltre che si trovi in una directory che il servente HTTP pubblica come '`http://`***nodo***`/cgi-bin/`' e che il servente HTTP sia anche disposto a eseguirlo in qualità di programma CGI, accedendo all'URI *http://**nodo**/cgi-bin/cgi-banale.sh* si dovrebbe osservare il risultato di questo programma. Se tutto si svolge presso l'elaboratore locale, l'URI diventa *http://localhost/cgi-bin/cgi-banale.sh*.

Figura 40.48. Risultato per l'utente della richiesta di accedere all'URI che punta allo script elementare ('**`cgi-banale.sh`**') che produce solo un output semplice senza interpretare alcun input.

Quando un cliente invia una richiesta di accedere a una risorsa che viene riconosciuta essere un programma *gateway*, il servente esegue questo programma e il suo standard output viene inviato in risposta al cliente, con l'aggiunta del codice di risultato iniziale: la preparazione del resto dell'intestazione è a carico del programma *gateway*. Quando il servente esegue il programma gli può inviare alcuni dati: in forma di argomenti della riga di comando, utilizzando le variabili di ambiente e anche attraverso lo standard input. Dipende dalla modalità della richiesta fatta dal cliente il modo con cui il programma *gateway* riceve i dati dal servente. È sufficiente realizzare uno script in grado di restituire tutti i dati che vengono forniti dal servente al programma *gateway* per comprendere il meccanismo (una copia di questo file dovrebbe essere disponibile anche qui: *allegati/cgi-test.sh*).

```
#!/bin/sh
#
# cgi-test.sh
#
echo "Content-type: text/html"
echo
echo "<HTML>"
```

```
echo "<HEAD>"
echo "<TITLE>Test CGI</TITLE>"
echo "</HEAD>"
echo "<BODY>\n";
echo "<H1>Test CGI</H1>"
echo "<PRE>"
echo "N. argomenti = $#"
echo "Argomenti    = $*"
echo
echo "SERVER_SOFTWARE = $SERVER_SOFTWARE"
echo "SERVER_NAME = $SERVER_NAME"
echo "GATEWAY_INTERFACE = $GATEWAY_INTERFACE"
echo "SERVER_PROTOCOL = $SERVER_PROTOCOL"
echo "SERVER_PORT = $SERVER_PORT"
echo "SERVER_ADMIN = $SERVER_ADMIN"
echo "REQUEST_METHOD = $REQUEST_METHOD"
echo "HTTP_ACCEPT = $HTTP_ACCEPT"
echo "HTTP_USER_AGENT = $HTTP_USER_AGENT"
echo "HTTP_CONNECTION = $HTTP_CONNECTION"
echo "PATH_INFO = $PATH_INFO"
echo "PATH_TRANSLATED = $PATH_TRANSLATED"
echo "SCRIPT_NAME = $SCRIPT_NAME"
echo "QUERY_STRING = $QUERY_STRING"
echo "REMOTE_HOST = $REMOTE_HOST"
echo "REMOTE_ADDR = $REMOTE_ADDR"
echo "REMOTE_USER = $REMOTE_USER"
echo "AUTH_TYPE = $AUTH_TYPE"
echo "CONTENT_TYPE = $CONTENT_TYPE"
echo "CONTENT_LENGTH = $CONTENT_LENGTH"
echo
echo "Standard input:"
cat
echo "</PRE>"
echo "</BODY>"
```

```
echo "</HTML>"
```

Figura 40.50. Richiamando lo script '**cgi-test.sh**' attraverso un URI, senza l'indicazione di alcuna stringa di richiesta, si ottiene lo stato delle variabili di ambiente fornite allo script stesso.

Eventualmente si può realizzare un altro programma, in Perl, che compie praticamente le stesse operazioni, ma in modo più preciso (una copia di questo file dovrebbe essere disponibile anche qui: *allegati/cgi-test.pl*).

```
#!/usr/bin/perl
#
# cgi-test.pl
#
print STDOUT ("Content-type: text/html\n");
print STDOUT ("\n");
print STDOUT ("<HTML>\n");
print STDOUT ("<HEAD>\n");
print STDOUT ("<TITLE>Test CGI</TITLE>\n");
print STDOUT ("</HEAD>\n");
print STDOUT ("<BODY>\n");
print STDOUT ("<H1>Test CGI</H1>\n");
print STDOUT ("<PRE>\n");
print STDOUT ("N. argomenti = $#ARGV\n");
print STDOUT ("Argomenti    = @ARGV\n");
print STDOUT ("\n");
#
foreach $var_amb (keys %ENV)
  {
    print STDOUT ("$var_amb = $ENV{$var_amb}\n");
  }
#
print STDOUT ("\n");
print STDOUT ("Standard input:");
#
while ($riga = <STDIN>)
  {
    print STDOUT ("$riga");
  }
#
print STDOUT ("</PRE>\n");
print STDOUT ("</BODY>\n");
print STDOUT ("</HTML>\n");
```

40.4.4 Percorso aggiuntivo

«

Esiste un metodo molto semplice per passare a un programma CGI un'informazione costituita da un percorso: quando si richiede un URI che punta a un programma CGI, ma seguito immediatamente e senza separazioni addizionali da un percorso che indichi un file o una directory, il programma CGI viene avviato e riceve questa informazione all'interno di una variabili di ambiente.

Per verificare come funzionano questi «percorsi aggiuntivi», basta usare lo script di verifica '`cgi-test.sh`' (oppure anche '`cgi-test.pl`'), mostrato in precedenza. Richiamando questo script, si può tentare di raggiungere un percorso che non esiste: supponendo di indicare l'URI *http://**nodo**/cgi-bin/cgi-test.sh/ciao/come/stai*, lo script riceve (e mostra) la variabile di ambiente ***PATH_INFO*** con il valore '`/ciao/come/stai`', mentre la variabile ***PATH_TRANSLATED*** contiene la (presunta) traduzione di quel percorso in un percorso reale, corrispondente probabilmente a '***document_root***`/ciao/come/stai`'. Sta poi al programma CGI sapere cosa farsene di questa informazione.

40.4.5 Elementi «FORM»

«

Gli elementi '**`FORM`**' servono a generare per l'utente dei «formulari», ovvero maschere di inserimento dati. L'input ottenuto in questo modo viene assemblato in coppie '***nome=valore***'. È poi compito del programma CGI disassemblare e interpretare tali informazioni.

I formulari degli elementi '**`FORM`**' vengono generati dal programma cliente (cioè dal navigatore) in base alle direttive incontrate all'interno di un documento HTML. Ciò significa che l'apparenza di que-

sti formulari può essere diversa a seconda del programma cliente utilizzato e del sistema operativo.

Il documento HTML contenente formulari di questo tipo, ovviamente, può essere stato predisposto nel servente come file normale, oppure può essere generato dinamicamente da un programma CGI.

```
<FORM …>
    …
    …
</FORM>
```

Un documento HTML può contenere più elementi '**FORM**', purché non siano annidati. L'elemento '**FORM**' può contenere degli attributi che ne definiscono il comportamento generale (ovviamente gli attributi si inseriscono nel marcatore di apertura), mentre all'interno della zona definita dall'elemento '**FORM**' si possono inserire altri elementi di vario genere, il cui scopo è quello di permettere all'utente un tipo particolare di interazione.

40.4.5.1 Attributo «ACTION»

L'attributo '**ACTION**' dell'elemento '**FORM**' specifica l'URI a cui inviare i dati inseriti attraverso il formulario. Deve trattarsi evidentemente dell'indirizzo di un programma CGI in grado di gestirli. Intuitivamente si comprende che questo attributo non può mancare. L'esempio seguente mostra in che modo si possa inserire questo attributo.

```
<FORM ACTION="http://www.brot.dg/cgi-bin/mio_programma.pl" …>
```

40.4.5.2 Attributo «METHOD»

L'attributo '**METHOD**' dell'elemento '**FORM**' specifica il ***metodo*** della richiesta che deve essere fatta dal cliente. Utilizzando un elemento '**FORM**' sono disponibili due tipi: '**GET**' e '**POST**'. L'esempio seguente mostra una situazione in cui si definisce l'utilizzo del metodo '**POST**'.

```
<FORM ACTION="http://www.brot.dg/cgi-bin/mio_programma.pl" METHOD="POST">
```

40.4.6 Elementi dell'ambiente «FORM»

All'interno dell'ambiente delineato dall'elemento '**FORM**', cioè della zona delimitata dai marcatori '**<FORM>**' e '**</FORM>**', si può collocare sia testo normale, sia elementi specifici di questo ambiente. È stato ripetuto più volte che i dati inseriti attraverso questi elementi vengono assemblati in coppie '***nome=valore***'. Quello che manca da sapere è che tali coppie vengono unite successivamente attraverso il simbolo e-commerciale ('**&**'). Gli esempi proposti più avanti mostrano meglio questo comportamento.

Esistono pochi tipi di elementi atti a permettere l'input all'interno dell'ambiente dell'elemento '**FORM**'. Questi cambiano il loro comportamento e l'apparenza a seconda degli attributi che gli vengono indicati. Il tipo di elemento più comune è '**INPUT**':

```
<INPUT NAME=… TYPE=… …>
```

Tutti gli elementi che permettono l'input hanno in comune l'attributo '**NAME**' che è obbligatorio. Le sezioni seguenti mostrano alcuni degli elementi utilizzabili in un formulario.

40.4.6.1 INPUT generico

Si tratta di un elemento che consente l'inserimento di testo normale su una sola riga. Questo elemento non richiede l'indicazione del tipo, attraverso l'attributo '**`TYPE`**'.

Attributo	Descrizione
`size="`***n***`"`	Permette di definire la dimensione in caratteri del campo che si vuole visualizzare.
`maxlength="`***n***`"`	Permette di stabilire un limite massimo alla dimensione, in caratteri, del testo che si può immettere.
`value="`***x***`"`	Permette di definire un valore predefinito che appaia già all'interno del campo.

L'esempio seguente visualizza un campo di 20 caratteri all'interno del quale l'utente deve scrivere il nome di un colore. Nel campo appare già la scritta '**`giallo`**' che può essere modificata o cancellata a piacimento.

```
Inserisci il colore:
<INPUT NAME="colore" SIZE="20" VALUE="giallo">
```

40.4.6.2 INPUT type="password"

Si tratta di un elemento che consente la scrittura di testo normale nascondendone l'inserimento, come avviene di solito quando si introducono le parole d'ordine. Dal momento che, a parte l'oscuramento dell'input, il funzionamento è uguale a quello dei campi di input normali, si possono utilizzare anche gli stessi tipi di attributi. L'esempio seguente visualizza un campo di 20 caratteri all'interno del quale l'utente deve inserire la parola d'ordine richiesta.

```
Inserisci la password: <INPUT TYPE="password" NAME="password-utente" SIZE="20">
```

40.4.6.3 INPUT type="checkbox"

«

Si tratta di un elemento che visualizza una casellina da barrare (casella di spunta). Queste caselline appaiono senza selezione in modo predefinito, a meno che venga utilizzato l'attributo '**`CHECKED`**'. Se la casellina risulta selezionata, viene generata la coppia '***nome=valore***' corrispondente, altrimenti no.

Attributo	Descrizione
`value="`***`x`***`"`	Permette di definire un valore (o una stringa) da restituire nel caso in cui la casellina sia selezionata. Questo attributo è essenziale.
`checked="checked"`	Questo attributo vale in quanto presente o meno, assegnandovi l'unico valore possibile che corrisponde al nome dell'attributo stesso. Se viene inserito nell'elemento, la casellina risulta inizialmente selezionata.

L'esempio seguente visualizza una casellina già barrata inizialmente. Se viene lasciata così, selezionata, questo elemento genera la coppia '**`propaganda=SI`**'.

```
Barrare la casella se si desidera ricevere propaganda:
<INPUT TYPE="checkbox" NAME="propaganda" VALUE="SI"
 CHECKED="checked">
```

40.4.6.4 INPUT type="radio"

«

Si tratta di un elemento che permette la selezione esclusiva di un pulsante all'interno di un gruppo. In pratica, selezionandone uno, si deselezionano gli altri. Rispetto agli elementi visti in precedenza, questo richiede la presenza di più elementi dello stesso tipo, altri-

menti non ci sarebbe da scegliere. Il collegamento che stabilisce che i pulsanti appartengono allo stesso gruppo viene definito dal nome che rimane uguale.

Attributo	Descrizione
`value="`***x***`"`	Permette di definire un valore (o una stringa) da restituire nel caso in cui il bottone risulti selezionato. Questo attributo è essenziale.
`checked="checked"`	Questo attributo vale in quanto presente o meno, assegnandovi l'unico valore possibile che corrisponde al nome dell'attributo stesso. Se viene inserito nell'elemento, il bottone risulta inizialmente selezionato.

L'esempio seguente visualizza tre pulsanti, di cui il primo già selezionato, per la scelta di un tipo di contenitore. I tre bottoni sono collegati insieme perché hanno lo stesso valore associato all'attributo '**`NAME`**'.

```
Selezionare il contenitore dell'elaboratore:
<INPUT TYPE="radio" NAME="contenitore" VALUE="orizzontale" CHECKED="checked">
<INPUT TYPE="radio" NAME="contenitore" VALUE="torre">
<INPUT TYPE="radio" NAME="contenitore" VALUE="minitorre">
```

40.4.6.5 INPUT type="submit"

Questo tipo di elemento visualizza un tasto contenente un'etichetta; selezionandolo si ottiene l'invio dei dati contenuti nel formulario in cui si trova. L'etichetta che appare sul pulsante in modo predefinito dipende dal cliente e potrebbe trattarsi di '**`Submit`**' o qualcosa del genere.

Questo elemento è diverso dagli altri in quanto non è previsto l'uso dell'attributo '**`NAME`**'. Infatti non viene generato alcun dato da

questo, ma solo l'invio dei dati contenuti nell'elemento '**FORM**'.

Attributo	Descrizione
`src="`***uri***`"`	Permette di indicare l'URI di un'immagine da utilizzare come pulsante.
`value="`***x***`"`	Permette di indicare un'etichetta alternativa a quella che verrebbe messa automaticamente dal programma cliente.

L'esempio seguente visualizza un tasto sul quale appare la scritta '**Invia la richiesta**'. Selezionandolo viene inviato il contenuto del formulario.

```
<INPUT TYPE="submit" VALUE="Invia la richiesta">
```

40.4.6.6 INPUT type="image"

«

Si tratta di una sorta di tasto di invio (*submit*) che in più aggiunge le coordinate in cui si trova il puntatore nel momento del clic. In un certo senso assomiglia anche agli elementi con l'attributo '**ISMAP**' descritto prima di affrontare gli elementi '**FORM**'.

Attributo	Descrizione
`src="`***uri***`"`	Permette di indicare l'URI dell'immagine da utilizzare come base. Questo attributo è obbligatorio data la natura dell'elemento.

L'esempio seguente visualizza l'immagine '`immagine.jpg`' e se viene fatto un clic con il puntatore del mouse sulla sua superficie, vengono inviati i dati del formulario, assieme anche alle coordinate relative all'immagine.

```
<INPUT TYPE="image" NAME="immagine" SRC="/immagine.jpg">
```

40.4.6.7 INPUT type="hidden"

Questo tipo di elemento, a prima vista, non ha alcun senso: permette di inserire dei campi nascosti, cosa che serve a generare una coppia '***nome=valore***' fissa.

È già stato chiarito che il protocollo HTTP non ha alcun controllo sullo stato delle transazioni, o meglio, ogni richiesta si conclude con una risposta. In questo modo, è compito del programma CGI mantenere il filo delle operazioni che si stanno svolgendo. Una delle tecniche con cui è possibile ottenere questo risultato è quella di restituire un formulario contenente le informazioni già inserite nelle fasi precedenti.

Ci sono anche altre situazioni in cui i dati nascosti e predefiniti sono utili, ma per il momento è sufficiente tenere a mente che esiste la possibilità.

Attributo	Descrizione
`value="`***x***`"`	Definisce il valore o la stringa nascosti. Tale argomento è obbligatorio per questo tipo di elemento.

L'esempio seguente fa in modo che il formulario contenga anche la coppia '**`nominativo=Tizio`**' che altrimenti, si suppone, renderebbe inutilizzabili gli altri dati inseriti dall'utente.

```
<INPUT TYPE="hidden" NAME="nominativo" VALUE="Tizio">
```

40.4.6.8 Elemento «TEXTAREA»

«

Questo elemento permette all'utente di inserire un testo su più righe. L'interruzione di riga, in questo caso, è fatta utilizzando la sequenza <*CR*><*LF*>. Questo particolare va tenuto presente in fase di programmazione, dal momento che gli ambienti Unix (in particolare i sistemi GNU) utilizzano l'interruzione di riga rappresentata con il solo carattere <*LF*>.

Attributo	Descrizione
`rows="`***n***`"`	Stabilisce il numero di righe dell'area di inserimento.
`cols="`***n***`"`	Stabilisce il numero di colonne dell'area di inserimento.

L'esempio seguente visualizza un'area per l'inserimento di testo su più righe. L'area visibile ha la dimensione di sette righe per 40 colonne e contiene già il testo '**`CIAO!`**' che può essere modificato o sostituito con qualcos'altro.

```
<TEXTAREA NAME="messaggio" ROWS="7" COLS="40" >
CIAO!
</TEXTAREA>
```

40.4.6.9 Elementi «SELECT» e «OPTION»

«

L'elemento '**`SELECT`**' delimita un ambiente attraverso cui si definiscono diverse scelte possibili, che normalmente appaiono in forma di menù a scomparsa. Per questo, oltre a '**`SELECT`**' si devono utilizzare degli elementi '**`OPTION`**' con cui si indicano tali scelte possibili. Va tenuto in considerazione che l'attributo '**`NAME`**' viene indicato nell'elemento '**`SELECT`**' (nel marcatore di apertura).

Attributo di SELECT	Descrizione
`multiple="multiple"`	Questo attributo vale in quanto presente o meno, assegnandovi l'unico valore possibile che corrisponde al nome dell'attributo stesso. Se presente, indica che sono ammissibili selezioni multiple, altrimenti è consentita la scelta di una sola voce.

Attributo di OPTION	Descrizione
`value="`***x***`"`	Definisce il valore (numero o stringa) da abbinare alla scelta eventuale. La stringa che appare all'utente è quella che segue il marcatore '**`OPTION`**' di apertura; se mancasse l'attributo '**`VALUE`**', sarebbe quella stessa stringa a essere restituita in abbinamento al nome definito nel marcatore '**`SELECT`**'.
`selected="selected"`	La presenza di questo attributo, a cui si assegna lo stesso nome dell'attributo, definisce una selezione predefinita.

L'esempio seguente presenta un menù di scelta a scomparsa per la selezione di un colore che poi viene convertito in un codice numerico corrispondente. Il nero, corrispondente allo zero, risulta predefinito.

```
<SELECT NAME="codice-colori">
    <OPTION VALUE="0" SELECTED="selected">Nero
    <OPTION VALUE="1">Marrone
    <OPTION VALUE="2">Rosso
    <OPTION VALUE="3">Arancio
    <OPTION VALUE="4">Giallo
    <OPTION VALUE="5">Verde
    <OPTION VALUE="6">Blu
    <OPTION VALUE="7">Viola
    <OPTION VALUE="8">Grigio
    <OPTION VALUE="9">Bianco
</SELECT>
```

40.4.7 Metodi e variabili

«

Esistono differenze nel modo con cui i programmi CGI ricevono le informazioni dal servente. Il modo fondamentale attraverso cui ciò viene controllato dal programma cliente è la scelta del ***metodo*** della richiesta: '**`GET`**' o '**`POST`**'. Fino a questo punto sono stati visti esempi che utilizzano esclusivamente il metodo '**`GET`**'.

Quando un programma cliente invia una richiesta utilizzando il metodo '**`GET`**' appende all'URI tutte le informazioni aggiuntive necessarie. In pratica, l'URI stesso comprende l'informazione. Per convenzione, la richiesta è distinta dalla parte dell'URI che identifica la risorsa attraverso un punto interrogativo, come nell'esempio seguente, dove la parola '**`ciao`**' è l'informazione aggiuntiva che rappresenta l'input per il programma '**`cgi-test.sh`**':

http://www.brot.dg/cgi-bin/cgi-test.sh?ciao

Il programma CGI riceve la «richiesta», inviata attraverso il metodo '**`GET`**', nella variabile di ambiente '**`QUERY_STRING`**'.

```
http://www.brot.dg/cgi-bin/cgi-test.sh?nome=Pinco&cognome=Pallino&sesso=M
```

L'URI mostrato sopra rappresenta una richiesta proveniente (presumibilmente) da un formulario HTML, per la presenza dei simboli di assegnamento. Come si può osservare, ogni coppia '***nome=valore***' è collegata alla successiva attraverso il simbolo e-commerciale ('**`&`**'). Il metodo '**`GET`**', in quanto aggiunge all'URI la stringa di richiesta, permette all'utente di controllare e di memorizzare il flusso di dati, per esempio attraverso un segnalibro (*bookmark*). In pratica, con la semplice memorizzazione dell'URI, l'utente può riprendere un'operazione di inserimento di dati, senza dover ricominciare tutto dall'inizio. Lo svantaggio nell'utilizzo di tale metodo sta nel fatto che esiste un limite alla dimensione degli URI e di conseguenza anche alla quantità di dati che gli si possono accodare.

Il metodo '**`POST`**' è stato progettato per porre rimedio ai limiti dell'altro metodo. Con questo, i dati dei formulari HTML vengono inviati in modo separato dall'URI, mentre il programma CGI li riceve dal programma servente attraverso lo standard input (invece che dalla variabile di ambiente ***QUERY_STRING***). Sotto questo aspetto, il metodo '**`POST`**' è generalmente preferibile.[4]

Le informazioni recepite da un programma CGI non si limitano alla «richiesta», giunta attraverso la variabile ***QUERY_STRING*** oppure dallo standard input: altre variabili di ambiente sono importanti per completare il contesto di lavoro.

Tabella 40.73. Variabili di ambiente contenenti informazioni sul servente.

Variabile	Descrizione
`SERVER_SOFTWARE`	Il nome e la versione del software utilizzato come servente.
`SERVER_NAME`	Il nome del servente.
`SERVER_PROTOCOL`	Il nome e la versione del protocollo utilizzato dal servente.
`SERVER_PORT`	Il numero della porta di comunicazione utilizzata dal servente.
`GATEWAY_INTERFACE`	Letteralmente, è l'interfaccia *gateway*, ovvero la versione del protocollo CGI utilizzato dal servente.
`PATH_INFO`	Quando l'URI contiene l'indicazione di un percorso aggiuntivo, questa variabile riceve quel percorso.
`PATH_TRANSLATED`	Questa variabile viene utilizzata assieme a ***PATH_INFO***, per indicare il percorso reale nel file system che ospita il servente.
`SCRIPT_NAME`	La parte dell'URI che identifica il percorso del programma CGI utilizzato.

Tabella 40.74. Variabili di ambiente contenenti informazioni sulla connessione cliente-servente.

Variabile	Descrizione
`REQUEST_METHOD`	Il metodo della richiesta ('**`GET`**', '**`POST`**').
`REMOTE_HOST`	Il nome del cliente. Se il nome non è disponibile, si deve fare uso della variabile ***REMOTE_ADDR*** che contiene l'indirizzo IP.
`REMOTE_ADDR`	Indirizzo IP del cliente.

Variabile	Descrizione
`AUTH_TYPE`	Contiene l'eventuale metodo di autenticazione.
`REMOTE_USER`	Il nome dell'utente se si utilizza l'autenticazione.

Tabella 40.75. Variabili di ambiente contenenti informazioni passate dal cliente al servente.

Variabile	Descrizione
`QUERY_STRING`	Contiene la stringa di richiesta se si utilizza il metodo '**`GET`**'.
`CONTENT_LENGTH`	Contiene la dimensione in byte (ottetti) dei dati ricevuti dal cliente. Questa informazione è disponibile solo se si utilizza il metodo '**`POST`**'.
`CONTENT_TYPE`	Contiene la definizione del tipo di codifica dei dati ricevuti dal cliente e riguarda solo il metodo '**`POST`**'. La codifica più comune è '**`application/x-www-form-urlencoded`**' e significa che i dati sono stati codificati secondo lo standard utilizzato per il metodo '**`GET`**': gli spazi sono convertiti in '**`+`**' e tutti i simboli speciali secondo la forma '**`%`*`hh`***', dove ***`hh`*** sono due cifre esadecimali.

Quando il cliente invia una richiesta al servente, prepara un'intestazione all'interno della quale possono essere inseriti diversi campi. Il contenuto di questi campi viene tradotto in altrettante variabili di ambiente il cui nome inizia per '**`HTTP_`**' seguito dal nome del campo stesso. In particolare, i caratteri minuscoli sono convertiti in maiuscoli e i trattini normali sono sostituiti dal trattino basso. Segue la

descrizione di alcune di queste variabili.

Informazioni addizionali dal cliente.

Variabile	Descrizione
`HTTP_ACCEPT`	Equivale al campo '**`Accept`**'.
`HTTP_USER_AGENT`	Equivale al campo '**`User-Agent`**'.

40.4.7.1 Un po' di pratica

«

Prima di iniziare a pensare a dei programmi CGI concludenti, conviene verificare quanto scritto attraverso i programmi di analisi mostrati in precedenza: '**`cgi-test.sh`**' oppure '**`cgi-test.pl`**'. Negli esempi viene mostrato sempre il primo dei due, anche se il migliore per queste cose sarebbe il secondo.

Si può realizzare una pagina HTML contenente dei formulari, come nell'esempio seguente.[5] Una copia di questo file dovrebbe essere disponibile all'indirizzo *allegati/form-test.html* ; inoltre, l'immagine costituita dal file '`test.jpg`' dovrebbe essere disponibile da *allegati/test.jpg* .

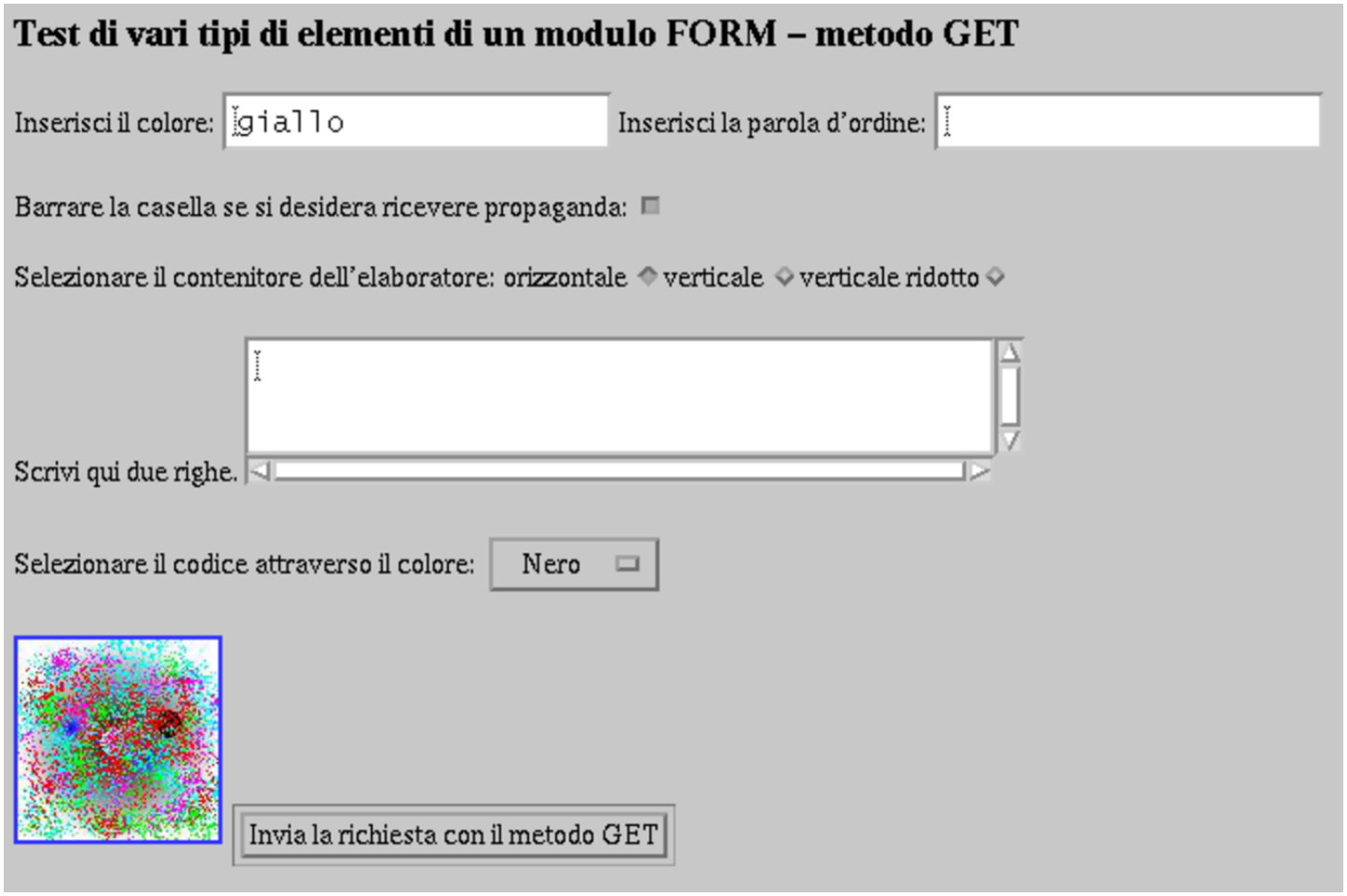
```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
        "http://www.w3.org/TR/html4/strict.dtd">
<!-- form-test.html -->
<HTML>
<HEAD>
    <TITLE>Verifica del funzionamento dei FORM</TITLE>
</HEAD>
<BODY>
    <H2>Test di vari tipi di elementi di un modulo FORM - metodo GET</H2>
    <FORM ACTION="/cgi-bin/cgi-test.sh" METHOD="GET">
        <P><INPUT TYPE="hidden" NAME="nominativo" VALUE="Tizio">
        Inserisci il colore:
            <INPUT NAME="colore" SIZE="20" VALUE="giallo">
```

```
        Inserisci la parola d'ordine:
            <INPUT TYPE="password" NAME="password-utente" SIZE="20">
        <P>Barrare la casella se si desidera ricevere propaganda:
            <INPUT TYPE="checkbox" NAME="propaganda" VALUE="SI"
                CHECKED="checked">
        <P>Selezionare il contenitore dell'elaboratore:
            orizzontale <INPUT TYPE="radio" NAME="case"
                VALUE="desktop" CHECKED="checked">
            verticale <INPUT TYPE="radio" NAME="case"
                VALUE="tower">
            verticale ridotto<INPUT TYPE="radio" NAME="case"
                VALUE="minitower">
        <P>Scrivi qui due righe.
            <TEXTAREA NAME="messaggio" ROWS="3" COLS="40"></TEXTAREA>
        <P>Selezionare il codice attraverso il colore:
            <SELECT NAME="codice-colori">
                <OPTION VALUE="0" SELECTED="selected">Nero
                <OPTION VALUE="1">Marrone
                <OPTION VALUE="2">Rosso
                <OPTION VALUE="3">Arancio
                <OPTION VALUE="4">Giallo
                <OPTION VALUE="5">Verde
                <OPTION VALUE="6">Blu
                <OPTION VALUE="7">Viola
                <OPTION VALUE="8">Grigio
                <OPTION VALUE="9">Bianco
            </SELECT>
        <INPUT TYPE="image" NAME="immagine" SRC="/test.jpg">
        <INPUT TYPE="submit" VALUE="Invia la richiesta con il metodo GET">
    </FORM>
    <HR>
    <H2>Test di vari tipi di elementi di un modulo FORM - metodo POST</H2>
    <FORM ACTION="/cgi-bin/cgi-test.sh" METHOD="POST">
        <P><INPUT TYPE="hidden" NAME="nominativo" VALUE="Tizio">
        Inserisci il colore:
            <INPUT NAME="colore" SIZE="20" VALUE="giallo">
        Inserisci la parola d'ordine:
            <INPUT TYPE="password" NAME="password-utente" SIZE="20">
        <P>Barrare la casella se si desidera ricevere propaganda:
            <INPUT TYPE="checkbox" NAME="propaganda" VALUE="SI"
                CHECKED="checked">
```

```
        <P>Selezionare il contenitore dell'elaboratore:
            orizzontale <INPUT TYPE="radio" NAME="case"
                VALUE="desktop" CHECKED="checked">
            verticale <INPUT TYPE="radio" NAME="case"
                VALUE="tower">
            verticale ridotto<INPUT TYPE="radio" NAME="case"
                VALUE="minitower">
        <P>Scrivi qui due righe.
            <TEXTAREA NAME="messaggio" ROWS="3" COLS="40"></TEXTAREA></P>
        <P>Selezionare il codice attraverso il colore:
            <SELECT NAME="codice-colori">
                <OPTION VALUE="0" SELECTED="selected">Nero
                <OPTION VALUE="1">Marrone
                <OPTION VALUE="2">Rosso
                <OPTION VALUE="3">Arancio
                <OPTION VALUE="4">Giallo
                <OPTION VALUE="5">Verde
                <OPTION VALUE="6">Blu
                <OPTION VALUE="7">Viola
                <OPTION VALUE="8">Grigio
                <OPTION VALUE="9">Bianco
            </SELECT>
        <INPUT TYPE="image" NAME="immagine" SRC="/test.jpg">
        <INPUT TYPE="submit" VALUE="Invia la richiesta con il metodo POST">
    </FORM>
</BODY>
</HTML>
```

Come si può vedere sono presenti due elementi '**FORM**' indipendenti: il primo utilizza il metodo '**GET**', il secondo invece il metodo '**POST**'. Entrambi gli elementi '**FORM**' richiamano il programma CGI '`/cgi-bin/cgi-test.sh`'.

Figura 40.78. Richiamando il file HTML dell'esempio, '`form-test.html`', con un programma cliente, si ottiene un formulario simile a quello di questa figura. Qui viene mostrata solo la prima parte, perché ciò che resta è la ripetizione dello stesso formulario utilizzando il metodo '**`POST`**'.



Si può già provare così, anche senza modificare alcunché. Se si invia la richiesta attraverso il formulario che utilizza il metodo '**`GET`**', si può osservare che la richiesta va a fare parte dell'URI del programma CGI; di conseguenza viene inserita nella variabile ***QUERY_STRING***. Altrimenti, con il metodo '**`POST`**' la richiesta si ottiene solo dallo standard input. In entrambi i casi, dovrebbe risultare codificata nello stesso modo (codifica URI).

```
nominativo=Tizio&colore=giallo&password-utente=&↩
↪propaganda=SI&case=desktop&messaggio=&↩
↪codice-colori=0
```

Si può osservare in particolare la presenza della coppia '**nominativo=Tizio**', inserita a titolo di esempio come campo nascosto e costante. Se invece di inviare il formulario attraverso la selezione del pulsante ('**submit**') si utilizza l'immagine, si ottiene una stringa simile a quella seguente:

```
nominativo=Tizio&colore=giallo&password-utente=&↩
↪propaganda=SI&case=desktop&messaggio=&↩
↪codice-colori=0&immagine.x=60&immagine.y=28
```

A questo punto, il lettore dovrebbe provare per conto proprio a compilare i campi, a modificare le selezioni, in modo da prendere dimestichezza con l'effetto generato dagli elementi '**FORM**'.

40.5 Programmazione CGI

«

Si introduce qui la programmazione per la realizzazione di programmi CGI in Perl. Il primo problema che si incontra quando si realizzano programmi del genere è l'analisi delle stringhe di richiesta, per arrivare alla loro scomposizione in modo da poterne gestire i dati. Per questo si utilizzano frequentemente librerie già pronte e ben collaudate, ma qui si vuole mostrare come lavorare partendo da zero.

Va osservato che negli esempi si usano prevalentemente delle richieste attraverso formulari HTML che utilizzano il metodo '**POST**'. Un buon programma CGI, tuttavia, dovrebbe essere in grado di gestire, indifferentemente, richieste fatte con i metodi '**GET**' e '**POST**'. Pertanto, queste spiegazioni non esauriscono l'argomento della programmazione CGI, ma affrontano solo alcuni dei suoi problemi.

Per una programmazione CGI efficace è consigliabile lo studio del linguaggio PHP (*http://www.php.net*).

40.5.1 Problemi

Prima di iniziare a realizzare programmi CGI, occorre fare mente locale alla situazione in cui si trova il programma, specialmente per la verifica del funzionamento dello stesso. Il programma viene eseguito attraverso una forma di intermediazione: è il servente HTTP a metterlo in funzione ed è sempre il servente a ricevere l'output che poi viene restituito al programma cliente.

In questa situazione, lo standard error del programma viene perduto, assieme alle eventuali segnalazioni di errore di qualunque tipo.

Prima di provare il funzionamento di un programma del genere, per quanto banale sia, occorre averlo analizzato sintatticamente attraverso gli strumenti che mette a disposizione il compilatore o l'interprete. L'utilizzo di Perl come linguaggio di programmazione, non richiedendo una fase di compilazione, tende a fare dimenticare che è necessaria un'analisi sintattica. Se non si verifica il programma, magari solo per un punto e virgola fuori posto, ci si trova di fronte al solito messaggio: «500 Errore interno del servente».

Nello stesso modo, sarebbe bene che il programma che si realizza sia in grado di funzionare in qualche modo anche al di fuori dell'ambiente creato dal servente HTTP.

È il caso di ricordare che il controllo sintattico di un programma Perl si ottiene nel modo seguente:

```
perl -c programma_perl
```

oppure ancora meglio con:

```
perl -c -w programma_perl
```

40.5.2 Decodifica

«

Si è accennato al fatto che un programma CGI non può fare a meno di occuparsi della decodifica delle stringhe di richiesta. Questo problema si scompone almeno nelle fasi seguenti:

- la suddivisione delle coppie '***nome=valore***';
- la separazione delle coppie;
- la decodifica URI.

I dati provenienti da un formulario HTML sono uniti assieme attraverso l'uso del simbolo e-commerciale ('**&**'). Per suddividerli si può creare un array dei vari elementi utilizzando la funzione '**split**'

```
@coppia = split ('&', $richiesta);
```

Le coppie '***nome=valore***' sono stringhe unite assieme attraverso il simbolo di assegnamento ('**=**'). La suddivisione avviene agevolmente attraverso la scomposizione in un array di due soli elementi. Solitamente si utilizza la scorciatoia seguente:

```
($nome, $valore) = split ('=', $coppia[$i]);
```

In pratica, si scompone il contenuto di un elemento dell'array '**@coppia**', visto nella sezione precedente.

La decodifica URI si scompone di due fasi:

- sostituzione del simbolo '**+**' con lo spazio;
- sostituzione dei codici '***%hh***' con il carattere corrispondente.

```
$valore   =~ tr/+/ /;

$nome   =~ s/%([A-Fa-f0-9][A-Fa-f0-9])/pack('c',hex($1))/ge;
$valore =~ s/%([A-Fa-f0-9][A-Fa-f0-9])/pack('c',hex($1))/ge;
```

Quello che segue è un esempio molto semplificato di due subroutine in grado, rispettivamente, di estrapolare le informazioni da una richiesta in modalità '**GET**' e in modalità '**POST**'. Le due subroutine restituiscono un hash (l'array associativo di Perl) corrispondente alle coppie di dati. Una copia di questo script dovrebbe essere disponibile anche qui: *allegati/mini-lib.pl* .

```
##
## mini-lib.pl
## Routine Perl utilizzabili da un programma CGI.
##
#
# &Decodifica_GET ()
# Decodifica il contenuto della variabile $QUERY_STRING e lo
# restituisce in un hash.
#
sub Decodifica_GET
{
    local ($richiesta) = $ENV{'QUERY_STRING'};
    #
    local (@coppia)    = ();
    local ($elemento)  = "";
    local ($nome)      = "";
    local ($valore)    = "";
    local (%DATI)      = ();
```

```
    #
    # Suddivide la richiesta in un array di coppie
    # «nome=valore».
    @coppia = split ('&', $richiesta);
    #
    # Elabora ogni coppia contenuta nell'array.
    foreach $elemento (@coppia)
      {
        #
        # Scompone la coppia.
        ($nome, $valore) = split ('=', $elemento);
        #
        # Trasforma «+» in spazio.
        $valore =~ tr/+/ /;
        #
        # Trasforma «%hh» nel carattere corrispondente.
        $nome
         =~ s/%([A-Fa-f0-9][A-Fa-f0-9])/pack('c',hex($1))/ge;
        $valore
         =~ s/%([A-Fa-f0-9][A-Fa-f0-9])/pack('c',hex($1))/ge;
        #
        # Aggiunge la coppia decodificata in un hash.
        $DATI{$nome} = $valore;
      }
    #
    # Restituisce l'hash delle coppie ( nome => valore ).
    return (%DATI);
}
#
# &Decodifica_POST ()
# Decodifica quanto proveniente dallo standard input e lo
# restituisce in un hash.
sub Decodifica_POST
{
    local ($richiesta) = "";
```

```
    #
    local (@coppia)    = ();
    local ($elemento)  = "";
    local ($nome)      = "";
    local ($valore)    = "";
    local (%DATI)      = ();
    #
    # Legge lo standard input.
    read (STDIN, $richiesta, $ENV{CONTENT_LENGTH});
    #
    # Suddivide la richiesta in un array di coppie
    # «nome=valore».
    @coppia = split ('&', $richiesta);
    #
    # Elabora ogni coppia contenuta nell'array.
    foreach $elemento (@coppia)
      {
        #
        # Scompone la coppia.
        ($nome, $valore) = split ('=', $elemento);
        #
        # Trasforma «+» in spazio.
        $valore =~ tr/+/ /;
        #
        # Trasforma «%hh» nel carattere corrispondente.
        $nome
         =~ s/%([A-Fa-f0-9][A-Fa-f0-9])/pack('c',hex($1))/ge;
        $valore
         =~ s/%([A-Fa-f0-9][A-Fa-f0-9])/pack('c',hex($1))/ge;
        #
        # Aggiunge la coppia decodificata in un hash.
        $DATI{$nome} = $valore;
      }
    #
    # Restituisce l'hash delle coppie ( nome => valore ).
```

```
    return (%DATI);
}
#
# Trattandosi di una libreria, l'ultima riga deve restituire
# un valore equiparabile a TRUE.
1;
#
```

Un programma banale che potrebbe fare uso di questa libreria, è il seguente. Si occupa solo di restituire i dati ottenuti dall'hash contenente le coppie '***nome=>valore***'. Una copia di questo script dovrebbe essere disponibile anche qui: *allegati/form.pl* .

```
#!/usr/bin/perl
##
## form.pl
##
require ('mini-lib.pl');
print STDOUT ("Content-type: text/html\n");
print STDOUT ("\n");
print STDOUT ("<HTML>\n");
print STDOUT ("<HEAD>\n");
print STDOUT ("<TITLE>Metodo $ENV{REQUEST_METHOD}</TITLE>\n");
print STDOUT ("</HEAD>\n");
print STDOUT ("<BODY>\n");
print STDOUT ("<H1>Metodo $ENV{REQUEST_METHOD}</H1>\n");
print STDOUT ("<PRE>\n");
if ($ENV{REQUEST_METHOD} eq 'GET')
  {
    %DATI = &Decodifica_GET;
  }
elsif ($ENV{REQUEST_METHOD} eq 'POST')
  {
    %DATI = &Decodifica_POST;
  }
```

```
else
  {
    print STDOUT ("Il metodo della richiesta ");
    print STDOUT ("non è gestibile.\n");
  }
@nomi = keys (%DATI);
foreach $nome (@nomi)
  {
    print STDOUT ("$nome = $DATI{$nome}\n");
  }
print STDOUT ("</PRE>\n");
print STDOUT ("</BODY>\n");
print STDOUT ("</HTML>\n");
```

Il programma '**form.pl**', appena mostrato, incorpora inizialmente la libreria presentata prima, '**mini-lib.pl**', quindi, a seconda del metodo utilizzato per la richiesta, chiama la subroutine adatta. Al termine, restituisce semplicemente l'elenco dei dati ottenuti.

40.5.3 Esempi elementari di applicazioni CGI

Nelle sezioni seguenti si mostrano alcuni esempi elementari di applicazioni CGI. Si tratta dell'accesso pubblico alla documentazione interna di un sistema operativo Unix comune, attraverso '**apropos**', '**whatis**' e '**man**'.

Per questi tre tipi di interrogazioni si prepara un solo file HTML di partenza, contenente tre elementi '**FORM**' distinti, ognuno dei quali invia una richiesta a un diverso programma CGI specializzato.

40.5.3.1 File «manuali.html»

«

Segue il sorgente del file ‘`manuali.html`’ contenente i tre elementi ‘**`FORM`**’ necessari per richiamare i programmi CGI in grado di fornire documentazione interna. Una copia di questo file dovrebbe essere disponibile anche qui: *allegati/manuali.html* .

```
<!DOCTYPE HTML PUBLIC "ISO/IEC 15445:2000//DTD HTML//EN">
<!-- manuali.html -->
<HTML>
<HEAD>
    <TITLE>Manualistica</TITLE>
</HEAD>
<BODY>
<H1>Manualistica</H1>
    <FORM ACTION="/cgi-bin/apropos.pl" METHOD="GET">
        <P>apropos <INPUT NAME="apropos" SIZE="30">
        <INPUT TYPE="submit" VALUE="Invio">
    </FORM>
    <FORM ACTION="/cgi-bin/whatis.pl" METHOD="GET">
        <P>whatis <INPUT NAME="whatis" SIZE="30">
            <INPUT TYPE="submit" VALUE="Invio">
    </FORM>
    <FORM ACTION="/cgi-bin/man.pl" METHOD="GET">
        <P>man 
            <SELECT NAME="sezione">
                <OPTION VALUE="" SELECTED="selected">predefinito
                <OPTION VALUE="1">comandi utente
                <OPTION VALUE="2">chiamate di sistema
                <OPTION VALUE="3">chiamate di libreria
                <OPTION VALUE="4">dispositivi
                <OPTION VALUE="5">formati dei file
                <OPTION VALUE="6">giochi
                <OPTION VALUE="7">varie
                <OPTION VALUE="8">comandi di sistema
                <OPTION VALUE="9">routine del kernel
            </SELECT>
```

```
            <INPUT NAME="man" SIZE="30">
            <INPUT TYPE="submit" VALUE="Invio">
    </FORM>
</BODY>
</HTML>
```

Figura 40.87. Il formulario ‘`manuali.html`’.

Ognuno dei tre elementi ‘**`FORM`**’ permette di indicare una stringa da utilizzare per ottenere informazioni. Per ogni elementi ‘**`FORM`**’ c’è un proprio tasto di invio indipendente con il quale si decide implicitamente il tipo di informazione che si vuole avere: *apropos*, *whatis* o *man*. Dei tre tipi di formulario, quello della richiesta per i file delle pagine di manuale è un po’ diverso, dal momento che potrebbe essere necessario indicare la sezione.

40.5.3.2 File «apropos.pl»

Segue il sorgente del programma ‘**`apropos.pl`**’, che si occupa di interrogare il sistema attraverso il comando ‘**`apropos`**’ e di restituire un file HTML con la risposta. Una copia di questo script dovrebbe essere disponibile anche qui: *allegati/apropos.pl* .

```
#!/usr/bin/perl
##
## apropos.pl
##
# Incorpora la libreria di decodifica dei dati.
require ('mini-lib.pl');
#
# &Metodo_non_gestibile ()
sub Metodo_non_gestibile
{
    print STDOUT ("Content-type: text/html\n");
    print STDOUT ("\n");
    print STDOUT ("<HTML>\n");
    print STDOUT ("<HEAD>\n");
    print STDOUT ("<TITLE>Errore</TITLE>\n");
    print STDOUT ("</HEAD>\n");
    print STDOUT ("<BODY>\n");
    print STDOUT ("<H1>Metodo $ENV{REQUEST_METHOD} ");
    print STDOUT ("non gestibile.</H1>\n");
    print STDOUT ("</BODY>\n");
    print STDOUT ("</HTML>\n");
}
#
# Inizio del programma.
local (%DATI)     = ();
local ($risposta) = "";
#
# Decodifica i dati in funzione del tipo di metodo della
# richiesta.
if ($ENV{REQUEST_METHOD} eq 'GET')
  {
    %DATI = &Decodifica_GET;
  }
elsif ($ENV{REQUEST_METHOD} eq 'POST')
  {
```

```
    %DATI = &Decodifica_POST;
  }
else
  {
    &Metodo_non_gestibile;
  }
#
# Rinvia la richiesta a apropos e ne restituisce l'esito.
if (open (APROPOS, "apropos $DATI{apropos} |"))
  {
    print STDOUT ("Content-type: text/html\n");
    print STDOUT ("\n");
    print STDOUT ("<HTML>\n");
    print STDOUT ("<HEAD>\n");
    print STDOUT ("<TITLE>apropos $DATI{apropos}</TITLE>\n");
    print STDOUT ("</HEAD>\n");
    print STDOUT ("<BODY>\n");
    print STDOUT ("<H1>apropos $DATI{apropos}</H1>\n");
    print STDOUT ("<PRE>\n");
    while ($risposta = <APROPOS>)
      {
        print $risposta;
      }
    print STDOUT ("</PRE>\n");
    print STDOUT ("</BODY>\n");
    print STDOUT ("</HTML>\n");
  }
else
  {
    print STDOUT ("Content-type: text/html\n");
    print STDOUT ("\n");
    print STDOUT ("<HTML>\n");
    print STDOUT ("<HEAD>\n");
    print STDOUT ("<TITLE>Errore</TITLE>\n");
    print STDOUT ("</HEAD>\n");
```

```
    print STDOUT ("<BODY>\n");
    print STDOUT ("<H1>Errore</H1>\n");
    print STDOUT ("Si è manifestato un errore ");
    print STDOUT ("durante l'inoltro ");
    print STDOUT ("della richiesta.\n");
    print STDOUT ("</BODY>\n");
    print STDOUT ("</HTML>\n");
  }
1;
```

Il programma è molto semplice: interpreta la richiesta ottenuta e ne estrae solo il valore abbinato all'informazione '**apropos**'; quindi esegue il comando '**apropos**' leggendone l'output che viene restituito in una pagina HTML molto semplice. Il punto più delicato di questo programma sta quindi nell'istruzione seguente:

```
open (APROPOS, "apropos $DATI{apropos} |")
```

Con questa viene abbinato un flusso di file a un comando il cui standard output viene letto successivamente e riemesso all'interno di una pagina HTML con il ciclo seguente:

```
while ($risposta = <APROPOS>)
  {
   print STDOUT ($risposta);
  }
```

Figura 40.91. Il risultato di un'interrogazione *apropos* per la parola «manual».

apropos manual

```
man (1)              - format and display the on-line manual pages
perlxs (1)           - XS language reference manual
whereis (1)          - locate the binary, source, and manual page files for a command
xman (1)             - Manual page display program for the X Window System
```

40.5.3.3 File «whatis.pl»

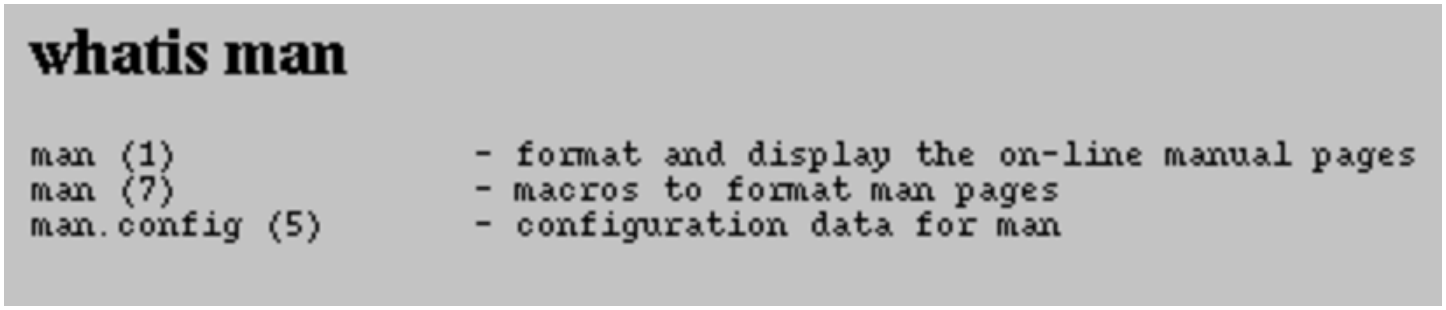
Segue il sorgente del programma '**whatis.pl**', che si occupa di interrogare il sistema attraverso il comando '**whatis**' e di restituire un file HTML con la risposta. È molto simile a '**apropos.pl**' appena mostrato, per cui qui alcune parti vengono tralasciate (in corrispondenza dei puntini di sospensione).

```
#!/usr/bin/perl
##
## whatis.pl
##
# Incorpora la libreria di decodifica dei dati.
require ('mini-lib.pl');
#
# &Metodo_non_gestibile ()
sub Metodo_non_gestibile
{
    ...
}
# Inizio del programma.
local (%DATI)     = ();
local ($risposta) = "";
#
# Decodifica i dati in funzione del tipo di metodo della
# richiesta.
if ($ENV{REQUEST_METHOD} eq 'GET')
  {
    %DATI = &Decodifica_GET;
  }
elsif ($ENV{REQUEST_METHOD} eq 'POST')
  {
    %DATI = &Decodifica_POST;
  }
else
```

```
  {
    &Metodo_non_gestibile;
  }
#
# Rinvia la richiesta a man e ne restituisce l'esito.
if (open( WHATIS, "whatis $DATI{whatis} |"))
  {
    print STDOUT ("Content-type: text/html\n");
    print STDOUT ("\n");
    print STDOUT ("<HTML>\n");
    print STDOUT ("<HEAD>\n");
    print STDOUT ("<TITLE>whatis $DATI{whatis}</TITLE>\n");
    print STDOUT ("</HEAD>\n");
    print STDOUT ("<BODY>\n");
    print STDOUT ("<H1>whatis $DATI{whatis}</H1>\n");
    print STDOUT ("<PRE>\n");
    while ($risposta = <WHATIS>)
      {
       print STDOUT ($risposta);
      }
    print STDOUT ("</PRE>\n");
    print STDOUT ("</BODY>\n");
    print STDOUT ("</HTML>\n");
  }
else
  {
    ...
  }
1;
```

Come si vede, si tratta della stessa cosa già vista nell'altro programma, con la differenza che la richiesta viene fatta al comando '**whatis**' invece che a '**apropos**'.

Figura 40.93. Il risultato di un'interrogazione *whatis* per la parola «man».



40.5.3.4 File «man.pl»

Segue il sorgente del programma '**man.pl**', che si occupa di interrogare il sistema operativo attraverso il comando '**man**' e di restituire un file HTML con la risposta. È molto simile agli altri due appena mostrati, per cui, anche in questo caso, alcune parti vengono tralasciate.

```
#!/usr/bin/perl
##
## man.pl
##
# Incorpora la libreria di decodifica dei dati.
require ('mini-lib.pl');
#
# &Metodo_non_gestibile ()
sub Metodo_non_gestibile
  {
    ...
  }
#
# Inizio del programma.
local (%DATI)     = ();
local ($risposta) = "";
#
# Decodifica i dati in funzione del tipo di metodo della
```

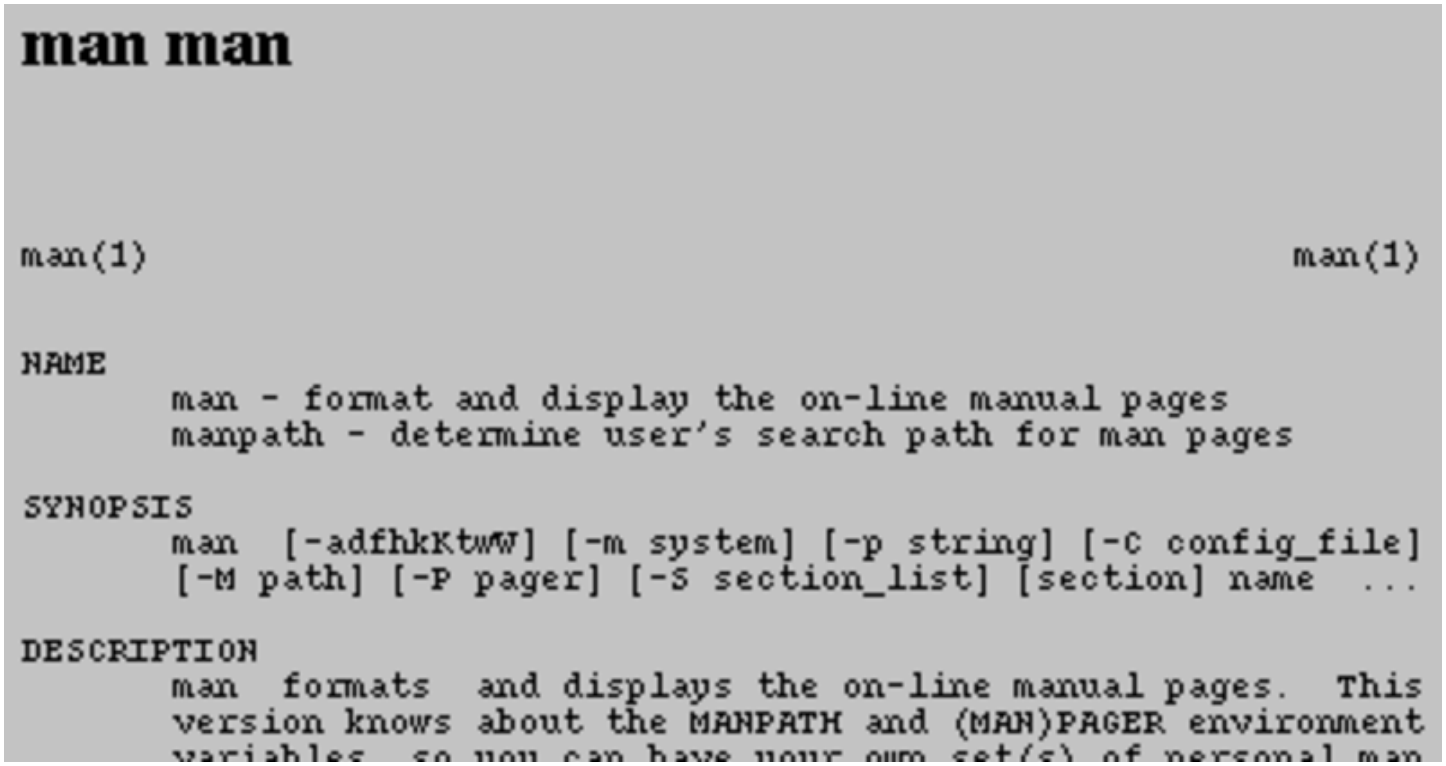
```
# richiesta.
if ($ENV{REQUEST_METHOD} eq 'GET')
  {
    %DATI = &Decodifica_GET;
  }
elsif ($ENV{REQUEST_METHOD} eq 'POST')
  {
    %DATI = &Decodifica_POST;
  }
else
  {
    &Metodo_non_gestibile;
  }
#
# Rinvia la richiesta a man e ne restituisce l'esito.
if (open (MAN, "man $DATI{sezione} $DATI{man} | col -bx |"))
  {
    print STDOUT ("Content-type: text/html\n");
    print STDOUT ("\n");
    print STDOUT ("<HTML>\n");
    print STDOUT ("<HEAD>\n");
    print STDOUT ("<TITLE>man $DATI{sezione} ");
    print STDOUT ("$DATI{man}</TITLE>\n");
    print STDOUT ("</HEAD>\n");
    print STDOUT ("<BODY>\n");
    print STDOUT ("<H1>man $DATI{sezione} ");
    print STDOUT ("$DATI{man}</H1>\n");
    print STDOUT ("<PRE>\n");
    while ($risposta = <MAN>)
      {
       print STDOUT ($risposta);
      }
    print STDOUT ("</PRE>\n");
    print STDOUT ("</BODY>\n");
```

```
    print STDOUT ("</HTML>\n");
  }
else
  {
    ...
  }
1;
```

La differenza fondamentale sta nel fatto che qui si utilizzano due informazioni: il nome del comando di cui si vuole ottenere la pagina di manuale e il numero della sezione. Un'altra cosa da osservare è il modo in cui è stato predisposto il comando: attraverso un condotto necessario a eliminare i caratteri di controllo che non potrebbero essere visualizzati nella pagina HTML.

```
open (MAN, "man $DATI{sezione} $DATI{man} | col -bx |")
```

Figura 40.96. Il risultato di un'interrogazione '**man**' per il comando '**man**', senza specificare la sezione.



40.5.4 Librerie CGI già pronte

Di solito, quando si parte da zero, conviene evitare di reinventarsi le subroutine necessarie a gestire i formulari HTML. Attraverso la rete si possono ottenere molti validi esempi già pronti e collaudati da più tempo.

Tra tutte, la libreria di subroutine Perl più diffusa per la gestione di formulari HTML sembra essere '**cgi-lib.pl**' di Steven Brenner.

40.6 Indicizzazione e motori di ricerca

Quando si imposta un servizio HTTP con molte informazioni utili ai visitatori, può essere importante mettere a disposizione un sistema di ricerca in base a delle parole chiave o delle stringhe più articolate.

Dove non ci si possa avvalere per questo di un servizio pubblico, occorre predisporne uno in proprio.

ht://Dig[6] è un motore di ricerca, vero e proprio, che ottiene i dati per la costruzione dei propri indici attraverso il protocollo HTTP. Pertanto, non si tratta di una scansione del file system pura e semplice.

L'installazione di ht://Dig richiede la preparazione di un file di configurazione, seguita immediatamente dalla preparazione di alcuni file, attraverso il programma '**`htdigconfig`**'; successivamente si passa alla scansione periodica degli indirizzi a cui si è interessati.

In generale, ht://Dig prevede una configurazione unica, in cui annotare tutti gli indirizzi da scandire, lasciando poi alla fase di ricerca l'onere di selezionare l'ambito del contesto cercato.

40.6.1 Configurazione e scansione periodica

La configurazione di ht://Dig si definisce in un file di testo normale (le righe bianche e quelle vuote vengono ignorate; i commenti sono preceduti dal simbolo '**#**'), rappresentato normalmente da '`/etc/htdig/htdig.conf`'. In generale, la directory che deve contenere il file di configurazione è stabilita in fase di compilazione dei sorgenti, mentre durante il funzionamento si possono indicare file di configurazione collocati altrove, ma solo in contesti particolari.

In ogni caso, secondo la filosofia di ht://Dig ci dovrebbe essere un solo file di configurazione, sotto il controllo dell'amministratore del sistema. Segue la descrizione di alcune direttive di questo file, che comunque viene fornito in modo predefinito con molti commenti esplicativi.

Direttiva	Descrizione
`database_dir:` ***directory***	Si stabilisce in questo modo la directory all'interno della quale devono essere inseriti i file che costituiscono la base di dati delle scansioni fatte da ht://Dig.
`start_url:` ***uri*** [***uri***] …	Permette di indicare uno o più indirizzi di partenza per le scansioni che si vogliono ottenere. Per esempio potrebbe trattarsi di indirizzi del tipo *http://**nodo**/* per scandire un sito intero, oppure *http://**nodo**/**percorso**/* per accedere soltanto a una porzione di questo.
`limit_urls_to: ${start_url}`	Questa direttiva serve a limitare la scansione a un certo ambito. Di solito si indicano gli stessi indirizzi usati nella direttiva '**`start_url`**', richiamandone il contenuto come si vede qui.
`exclude_urls:` ***modello*** [***modello***]	Consente di escludere dalla scansione tutti gli indirizzi che contengono una stringa tra quelle elencate in questa direttiva. Di solito si indicano stringhe del tipo '**`/cgi-bin/`**' e '**`.cgi`**', per impedire di accedere a programmi CGI.

Direttiva	Descrizione
`bad_extensions:` ***estensione*** [***estensione***]	Questa direttiva è simile a '**`exclude_urls`**', con la differenza che riguarda solo la parte finale di un indirizzo (l'estensione). Si indicano di solito tutte le estensioni che possono fare riferimento a file che ht://Dig non riesce ad analizzare.
`maintainer:` ***indirizzo_email***	Consente di specificare il responsabile della gestione del servizio.

Oltre al file '`/etc/htdig/htdig.conf`', ne esistono comunque degli altri, collocati sempre nella directory '`/etc/htdig/`', ma in generale non è necessario modificarli. Eventualmente, può essere conveniente in un secondo momento la traduzione dei file HTML di questa directory, dato che ht://Dig li usa quando costruisce le sue risposte mostrate attraverso un programma CGI apposito.

Alcuni di questi file contenuti nella directory '`/etc/htdig/`' servono per costruire una piccola base di dati iniziale che contiene informazioni su sinonini (generata dal file '`/etc/htdig/synonyms`') e sulle radici delle parole (generata dai file '`/etc/htdig/english.*`' e '`/etc/htdig/bad_words`'). Per questo si usa il programma '**`htdigconfig`**':

```
# htdigconfig [Invio]
```

Terminata questa fase iniziale, si passa alla scansione periodica di quanto programmato nella configurazione. Per questo si usa normalmente il programma '**`rundig`**' (potrebbe essere uno script che si av-

vale di altri programmi di ht://Dig, ma questo fatto non ha molta importanza). Conviene distinguere due possibilità:

1. # **rundig -a -i** [*Invio*]
2. # **rundig -a** [*Invio*]

Nel primo caso si tratta di una scansione in cui la base di dati precedente, se esiste, viene messa da parte senza cancellarla, ricostruendo comunque una base di dati nuova; nel secondo caso invece, la base di dati viene sì ricostruita, ma si tiene conto di quella precedente, aggiungendo soltanto le informazioni nuove e togliendo i riferimenti a file che non esistono più. Pertanto, conviene eseguire il primo comando con una periodicità che potrebbe essere settimanale, mentre il secondo va eseguito con una frequenza maggiore, anche giornaliera. Evidentemente, conviene usare per questo il sistema Cron.

È bene osservare che la scansione avviene attraverso il protocollo HTTP ed è possibile accumulare gli indici di un sito che si trova anche all'esterno del proprio elaboratore. Pertanto, quando si configura ht://Dig per raggiungere un elaboratore esterno, è bene considerare anche il traffico (il carico della rete) che l'aggiornamento degli indici può comportare.

Teoricamente, ht://Dig può indicizzare anche il contenuto di file PDF, PostScript e di altri formati, purché siano disponibili alcuni programmi di conversione. Tuttavia, non è conveniente abilitare questa funzionalità nella configurazione di ht://Dig, perché la scansione per l'accumulo delle informazioni diventa molto pesante, sia

per la rete, sia per l'elaborazione che ha luogo; inoltre, i visitatori che trovano le informazioni contenute in file di questo tipo, possono trovarsi poi in difficoltà, mentre è auspicabile che le stesse notizie siano accessibili anche attraverso pagine HTML normali. Pertanto, è bene prendere in considerazione la direttiva di configurazione '**`bad_extensions`**', aggiungendo tutte queste estensioni che non conviene prendere in considerazione.

40.6.2 Interrogazione del motore di ricerca

Il programma con il quale si interroga la base di dati costruita da ht://Dig è '**`htsearch`**', il quale si usa normalmente come programma CGI, ma si può utilizzare anche attraverso la riga di comando, tenendo conto però che la risposta è sempre in forma di pagina HTML. Data la sua natura, il programma viene installato normalmente all'interno della directory usata per i programmi CGI. Per esempio, potrebbe trattarsi dell'indirizzo *http://dinkel.brot.dg/cgi-bin/htsearch*. Segue la figura di ciò che si vede la prima volta (senza l'indicazione di una stringa di ricerca):

```
                    ht://Dig Search results
 ----------------------------------------------------------
No matches were found for ''

 Check the spelling of the search word(s) you used. If the
 spelling is correct and you only used one word, try using
 one or more similar search words with "Any."

 If the spelling is correct and you used more than one word
 with "Any," try using one or more similar search words with
 "Any."

 If the spelling is correct and you used more than one word
 with "All," try using one or more of the same words with
 "Any."
 ----------------------------------------------------------
 Match: [All____] Format: [Long_] Sort by: [Score________]
 Refine search: _____________________________ [ Search ]
 ----------------------------------------------------------
```

Nella parte finale della pagina si ottiene un formulario da compilare per la ricerca. Ecco cosa si può ottenere quando si indica qualche parola chiave significativa:

```
 ----------------------------------------------------------
   Documents 1 - 10 of 3811 matches. More *'s indicate a
   better match.
 ----------------------------------------------------------
Appunti di informatica libera * * * *
   ... ] [inizio] [fine] [indice generale] [violazione GPL]
   [licenze] [indice analitico] [volume] [parte] Capitolo
   259.  Convenzioni di «Appunti di informatica libera»
   Questo capitolo raccoglie alcune convenzioni importanti
   relative all'opera Appunti di informatica libera. Le
   annotazioni sulla terminologia ...
   http://dinkel.brot.dg/a2/prossima/HTML-2002.08/a2322.html
   08/19/02, 49757 bytes

Appunti di informatica libera * * * *
   ... ] [inizio] [fine] [indice generale] [violazione GPL]
   [licenze] [indice analitico] [volume] [parte] Capitolo
   259.  Convenzioni di «Appunti di informatica libera»
   Questo capitolo raccoglie alcune convenzioni importanti
   relative all'opera Appunti di informatica libera. Le
   annotazioni sulla terminologia ...
   http://dinkel.brot.dg/a2/dist/CD2/HTML/a2326.html
   07/21/02, 49503 bytes
```

Eventualmente, può essere conveniente realizzare un formulario HTML personalizzato, così da poter anche tradurre alcuni termini:[7]

```
<FORM METHOD="GET" ACTION="/cgi-bin/htsearch">
<P><INPUT TYPE="HIDDEN" NAME="config" VALUE="">
<INPUT TYPE="HIDDEN" NAME="restrict" VALUE="">
<INPUT TYPE="HIDDEN" NAME="exclude" VALUE="">
confronto:
<SELECT NAME="method">
    <OPTION VALUE="and" SELECTED="selected">di tutte le parole
    <OPTION VALUE="or">di almeno una parola
    <OPTION VALUE="boolean">booleano
```

```
</SELECT>

formato:
<SELECT NAME="format">
    <OPTION VALUE="builtin-long">lungo
    <OPTION VALUE="builtin-short">breve
</SELECT>

ordinato per:
<SELECT NAME="sort">
    <OPTION VALUE="score" SELECTED="selected">punteggio
    <OPTION VALUE="time">data
    <OPTION VALUE="title">titolo
    <OPTION VALUE="revscore">punteggio in modo inverso
    <OPTION VALUE="revtime">data in modo inverso
    <OPTION VALUE="revtitle">titolo in modo inverso
</SELECT>

<BR>
stringa di ricerca:
<INPUT TYPE="text" SIZE="40" NAME="words" VALUE="">
<INPUT TYPE="submit" VALUE="ricerca">

</FORM>
```

Attraverso la modifica di alcuni campi nascosti è possibile limitare la ricerca a un solo sito o a una porzione di questo. Per esempio, per richiedere una ricerca limitata esclusivamente a ciò che si articola a partire da *http://dinkel.brot.dg/a2/* (purché i dati relativi siano stati scanditi in precedenza), basta ritoccare la prima parte del formulario nel modo seguente:

```
<FORM METHOD="GET" ACTION="/cgi-bin/htsearch">
<P><INPUT TYPE="HIDDEN" NAME="config" VALUE="">
<INPUT TYPE="HIDDEN" NAME="restrict" VALUE="http://dinkel.brot.dg/a2/">
```

```
<INPUT TYPE="HIDDEN" NAME="exclude" VALUE="">
...
</FORM>
```

Inoltre, è possibile escludere espressamente qualcosa; per esempio si potrebbe voler ignorare quanto si articola sotto *http://dinkel.brot.dg/a2/pasticci/*:

```
<FORM METHOD="GET" ACTION="/cgi-bin/htsearch">
<P><INPUT TYPE="HIDDEN" NAME="config" VALUE="">
<INPUT TYPE="HIDDEN" NAME="restrict" VALUE="http://dinkel.brot.dg/a2/">
<INPUT TYPE="HIDDEN" NAME="exclude" VALUE="http://dinkel.brot.dg/a2/pasticci/">
...
</FORM>
```

È importante osservare che le stringhe di inclusione e quelle di esclusione vengono confrontate con una parte qualunque dell'indirizzo; per esempio è facile specificare delle estensioni, come in questo caso in cui si vogliono escludere i file che potrebbero essere in formato SGML:

```
<FORM METHOD="GET" ACTION="/cgi-bin/htsearch">
<P><INPUT TYPE="HIDDEN" NAME="config" VALUE="">
<INPUT TYPE="HIDDEN" NAME="restrict" VALUE="http://dinkel.brot.dg/a2/">
<INPUT TYPE="HIDDEN" NAME="exclude" VALUE=".sgml">
...
</FORM>
```

Quando si inseriscono delle limitazioni, come in questi esempi, le pagine che mostrano il risultato della ricerca aggiungono un formulario per altre ricerche, in cui valgono le stesse limitazioni di partenza.

Gli esempi mostrano tutti dei moduli che usano un metodo '**GET**' per accedere al programma CGI. ht://Dig funziona perfettamente anche con l'uso di un metodo POST, ma in tal modo viene a mancare la possibilità di memorizzare nei file delle registrazioni del servente HTTP interrogato l'indirizzo referente con la stringa di richiesta. In pratica, in tal modo, programmi come Webalizer non hanno poi la possibilità di estrapolare le interrogazioni fatte per raggiungere le pagine del sito a cui si riferiscono.

40.6.3 Configurazioni multiple

«

Anche se sconsigliabile secondo la filosofia di ht://Dig, è possibile gestire delle configurazioni multiple, ovvero più file di configurazione a cui si abbinano delle basi di dati differenti per gli indici. Tuttavia, è possibile collocare i file di configurazione alternativi solo nella stessa directory in cui è previsto quello normale, ovvero '`/etc/htdig/`', mantenendo l'estensione '`.conf`'. Per esempio, si può definire un file di configurazione alternativo, corrispondente a '`/etc/htdig/prova.conf`', mentre non si può usare il file '`/etc/htdig/prova.configura`'.

Una volta definita la configurazione alternativa, si deve procedere a generare la sua basi di dati con '**rundig**', aggiungendo l'opzione '**-c**', per esempio così:

```
# rundig -a -i -c /etc/htdig/prova.conf [Invio]
```

Successivamente, nel formulario usato per interrogare la basi di dati, si indica il riferimento alla configurazione '**prova**' (senza estensione e senza percorso):

```
<FORM METHOD="GET" ACTION="/cgi-bin/htsearch">
<P><INPUT TYPE="HIDDEN" NAME="config" VALUE="prova">
...
</FORM>
```

40.7 Statistiche di accesso

Dal momento che il protocollo HTTP è privo di stato, ogni operazione elementare inizia e conclude una connessione TCP, la quale può essere annotata nel file delle registrazioni del servente HTTP. Nella gestione di un sito che offre i suoi servizi attraverso il protocollo HTTP, può essere importante l'analisi dei file delle registrazioni del servente HTTP, per ottenere delle statistiche sugli accessi. L'analisi quotidiana di queste statistiche consente di capire meglio cosa cerca il pubblico e che tipo di reazione si ottiene a seguito di iniziative che fanno capo al proprio sito.[8]

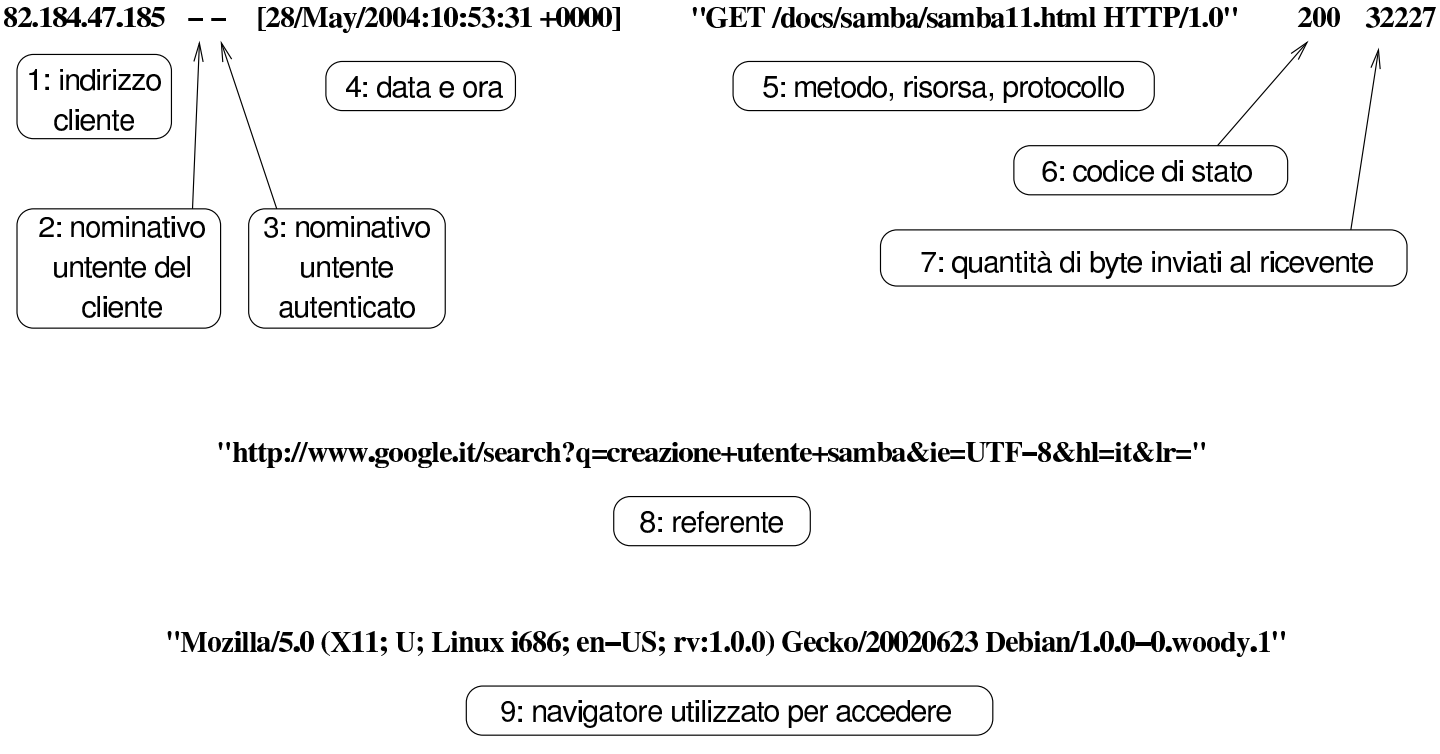
Fortunatamente, i serventi più comuni utilizzano delle annotazioni abbastanza compatibili. Il formato in questione standard per la registrazione degli accessi, viene definito *Common log format*, a cui si associa anche una variante più completa, definita come formato «combinato». In generale, se possibile, è meglio usare il formato combinato che contiene l'indicazione del referente, ovvero dell'indirizzo dal quale proviene il riferimento ipertestuale.

L'esempio seguente riguarda alcune righe di un registro di accesso organizzato secondo il formato combinato; si osservi che le righe appaiono spezzate per motivi tipografici:

```
82.184.47.185 - - [28/May/2004:10:53:31 +0000] "GET /docs/samba/samba11.html HTTP/1.0" 200 32227 ↩
↪"http://www.google.it/search?q=creazione+utente+samba&ie=UTF-8&hl=it&lr=" ↩
↪"Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.0.0) Gecko/20020623 Debian/1.0.0-0.woody.1"
82.184.47.185 - - [28/May/2004:10:53:32 +0000] "GET /docs/samba/7.jpg HTTP/1.0" 200 29924 ↩
↪"http://linuxdidattica.org/docs/samba/samba11.html" ↩
↪"Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.0.0) Gecko/20020623 Debian/1.0.0-0.woody.1"
82.184.47.185 - - [28/May/2004:10:53:33 +0000] "GET /docs/samba/8.jpg HTTP/1.0" 200 30174 ↩
↪"http://linuxdidattica.org/docs/samba/samba11.html" ↩
↪"Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.0.0) Gecko/20020623 Debian/1.0.0-0.woody.1"
82.184.47.185 - - [28/May/2004:10:53:33 +0000] "GET /docs/samba/6.jpg HTTP/1.0" 200 39877 ↩
↪"http://linuxdidattica.org/docs/samba/samba11.html" ↩
↪"Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.0.0) Gecko/20020623 Debian/1.0.0-0.woody.1"
82.184.47.185 - - [28/May/2004:10:53:34 +0000] "GET /docs/samba/9.jpg HTTP/1.0" 200 16244 ↩
↪"http://linuxdidattica.org/docs/samba/samba11.html" ↩
↪"Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.0.0) Gecko/20020623 Debian/1.0.0-0.woody.1"
82.184.47.185 - - [28/May/2004:10:53:34 +0000] "GET /docs/samba/10.jpg HTTP/1.0" 200 21050 ↩
↪"http://linuxdidattica.org/docs/samba/samba11.html" ↩
↪"Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.0.0) Gecko/20020623 Debian/1.0.0-0.woody.1"
82.184.47.185 - - [28/May/2004:10:53:35 +0000] "GET /docs/samba/11.jpg HTTP/1.0" 200 20936 ↩
↪"http://linuxdidattica.org/docs/samba/samba11.html" ↩
↪"Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.0.0) Gecko/20020623 Debian/1.0.0-0.woody.1"
```

Si comincia dalla prima riga per osservare che si tratta di un accesso con una richiesta secondo il metodo '**GET**', avente origine dall'indirizzo 82.184.47.185. Per la precisione, è stata prelevata la risorsa corrispondente a *http://**nodo**/docs/samba/samba11.html*. L'utente che ha richiesto questa risorsa lo ha fatto a partire da un riferimento abbastanza complesso, rappresentato verosimilmente da una pagina generata da un motore di ricerca, come si vede nella figura successiva.

Figura 40.106. Un record di un registro di accesso secondo il formato combinato. Si può osservare che in questo caso i campi numero due e numero tre non contengono informazioni. Il formato CLF tradizionale, a differenza di quello combinato, non ha gli ultimi due campi.



Continuando l'osservazione dell'esempio, si può vedere che a partire da *http://**nodo**/docs/samba/samba11.html* sono state raggiunte le risorse */docs/samba/7.jpg*, */docs/samba/8.jpg*, */docs/samba/6.jpg*, */docs/samba/9.jpg*, */docs/samba/10.jpg* e */docs/samba/11.jpg*, le quali sono evidentemente immagini inserite nella pagina di partenza.

L'informazione sull'indirizzo referente, ovvero sull'indirizzo di partenza, permette di comprendere l'importanza che può avere il riferimento fatto da qualcun altro verso le pagine del proprio sito. In altri termini, Tizio che indica nelle sue pagine un riferimento a un certo indirizzo esterno, fa una cortesia a quel sito, cosa che può essere valutata nel numero di accessi che in questo modo vi vengono

convogliati.

Tuttavia, le informazioni generate dal servente HTTP non sono sempre così dettagliate; spesso manca l'indicazione dell'indirizzo referente, a meno di richiedere espressamente tali notizie nella configurazione. L'esempio seguente riguarda una porzione della configurazione di Apache, in cui si dichiara il dominio virtuale *linuxdidattica.org* e gli si associa un file di registrazioni specifico ('`/var/log/apache/linuxdidattica.org-access.log`') con tutte le informazioni che Apache è in grado di dare:

```
<VirtualHost 62.152.34.13>
Servername linuxdidattica.org
DocumentRoot /home/www/linuxdidattica.org
CustomLog /var/log/apache/linuxdidattica.org-access.log full
</VirtualHost>
```

Il fatto di poter ottenere un file delle registrazioni separato per gli accessi a un dominio virtuale, oppure a un ramo del proprio sito, diventa importante, proprio per facilitare il lavoro successivo di lettura delle statistiche.

Eventualmente, se non è possibile ottenere dal servente HTTP un file delle registrazioni selettivo per un certo dominio virtuale, o per un certo ramo del proprio sito, si può intervenire con un programma realizzato appositamente per filtrare l'unico file a disposizione:

```
#!/usr/bin/perl
$modello = $ARGV[0];
$riga = "";
while ($riga = <STDIN>)
  {
    if ($riga =~ m{\"[A-Z]+ $modello.* HTTP/[0-9.]+\"})
      {
        print STDOUT ($riga);
      }
  }
```

Se questo programma viene chiamato '**filtra**' e il file delle registrazioni è '/var/log/httpd/access.log', per ottenere un file con gli accessi che si diramano a partire da *http://**nodo**/servizi/casa/*, si potrebbe usare il comando seguente:

```
# cat /var/log/httpd/access.log | filtra /servizi/casa/ ↩
↪> /var/log/tmp_servizi_casa.log[Invio]
```

In questo modo si creerebbe il file '/var/log/tmp_servizi_casa.log' con i soli record che interessano.

40.7.1 Webalizer

Webalizer[9] è un programma relativamente semplice per l'analisi di un file di registrazioni in formato CLF (*Common log format*) o in formato combinato, dal quale produce un rapporto statistico che può essere letto anche attraverso lo stesso servizio HTTP. In pratica, il rapporto che si ottiene è fatto di pagine HTML e di immagini contenenti i grafici dei vari rapporti statistici generati; queste pagine possono essere consultate localmente o a distanza, con un navigatore comune.

Webalizer si avvale di un solo file di configurazione che in condizio-

ni normali corrisponde a '`/etc/webalizer.conf`'. Tuttavia, nel file di configurazione si possono indicare espressamente il file delle registrazioni da analizzare e la directory di destinazione dei file delle statistiche; pertanto, se si gestiscono diversi siti virtuali, o comunque se quello che serve sono statistiche diverse in base al contesto di interesse, potrebbe essere conveniente la predisposizione di file di configurazione differenti, ognuno per l'obiettivo desiderato. Segue un elenco parziale delle direttive di questo file di configurazione, a cui si affianca l'opzione corrispondente dell'eseguibile '**`webalizer`**', quando disponibile.

Direttiva, opzione	Descrizione
`-c` ***file***	Permette di indicare un file di configurazione alternativo a quello predefinito.

`Debug yes|no`
 	
`-d`
	Permette di ottenere maggiori informazioni durante l'elaborazione delle statistiche.

`LogFile` ***file***	Permette di definire il file delle registrazioni da scandire.
`LogType [clf|ftp`↩ ↩`|squid]` `-F [clf|ftp|squid]`	Webalizer è in grado di analizzare file delle registrazioni in formati diversi, specificandolo con questa direttiva. Il formato corrispondente alla parola chiave '**`clf`**' è quello dei serventi HTTP comuni. La sigla '**`clf`**' sta per *Common log format* che però vale anche per il formato «combinato», ovvero quello che contiene le informazioni sul referente e sul tipo di navigatore.

Direttiva, opzione	Descrizione
`OutputDir` ***file*** `-o` ***file***	In questo modo si specifica la directory nella quale creare i file che compongono le statistiche.
`HostName` ***nome*** `-n` ***nome***	Permette di definire il nome del sito (reale o virtuale che sia) che viene inserito nei file delle statistiche.
`ReportTitle` ***nome*** `-t` ***nome***	Permette di modificare il titolo predefinito delle statistiche. Dopo il titolo si aggiunge il nome definito con la direttiva '**`HostName`**' o con l'opzione '**`-n`**'.
`VisitTimeout` ***n*** `-m` ***n***	Consente di stabilire il tempo di scadenza per la durata delle visite. In tal modo, un accesso proveniente dallo stesso indirizzo già visto più di ***n*** secondi prima, viene considerato una visita nuova e non semplicemente una richiesta di un accesso preesistente.
`PageType` ***modello*** `-P` ***modello***	Questa opzione che può essere usato più volte e consente di specificare l'estensione dei file da considerare come «pagine». Di solito si usa una stringa del tipo '**`htm*`**', per includere le pagine HTML comuni, ma può essere conveniente aggiungere anche altre estensioni, a seconda del modo in cui è organizzato il proprio sito.

Direttiva, opzione	Descrizione
`CountryGraph yes|no` `-Y yes|no`	Abilita o disabilita la visualizzazione del grafico delle nazionalità degli accessi, basato sulla parte finale del nome a dominio.
`DailyGraph yes|no`	Abilita o disabilita la visualizzazione del grafico giornaliero degli accessi.
`DailyStats yes|no`	Abilita o disabilita la visualizzazione della statistica giornaliera degli accessi.
`HourlyGraph yes|no` `-G yes|no`	Abilita o disabilita la visualizzazione del grafico orario degli accessi.
`HourlyStats yes|no` `-H yes|no`	Abilita o disabilita la visualizzazione della statistica oraria degli accessi.
`Incremental yes|no` `-p yes|no`	Abilitando questa opzione con la parola chiave '**`yes`**', si fa in modo che Webalizer tenga conto anche delle statistiche precedenti, in modo da non perdere dati quando il sistema di rotazione dei file delle registrazioni riparte con file vuoti.
`DNSCache` ***file***	Definisce il nome da dare a un file che Webalizer può usare per annotare degli indirizzi risolti in nomi. Questo file, assieme alla direttiva '**`DNSChildern`**', consente di ottenere i nomi delle origini degli accessi, quando è possibile risolverli.

Direttiva, opzione	Descrizione
`DNSChildren` ***n***	Assieme alla direttiva '**`DNSCache`**' abilita la risoluzione degli indirizzi in nomi a dominio, specificando il numero di processi elaborativi che devono occuparsi di questo lavoro.
`HideReferer` ***modello*** `-r` ***modello***	Fa in modo che nel resoconto dei referenti, non appaiano i nomi che corrispondono al modello.
`IgnoreReferer` ***modello***	Fa in modo che i record contenenti dei referenti corrispondenti al modello indicato vengano ignorati completamente.
`HideSite` ***modello*** `-s` ***modello***	Fa in modo che nel resoconto dell'origine degli accessi, non appaiano i nomi che corrispondono al modello.
`IgnoreSite` ***modello***	Fa in modo che i record contenenti origini corrispondenti al modello indicato vengano ignorati completamente.
`HideURL` ***modello*** `-u` ***modello***	Fa in modo che nel resoconto delle risorse richieste non appaiano i nomi che corrispondono al modello.
`IgnoreURL` ***modello***	Fa in modo che i record contenenti la richiesta di una risorsa corrispondente al modello indicato vengano ignorati completamente.

Direttiva, opzione	Descrizione

`AllSites yes|no`
`AllURLs yes|no`
`AllReferrers yes|no`
`AllAgents yes|no`
`AllSearchStr yes|no`
`AllUsers yes|no`
	Queste direttive, se attivate, fanno sì che rimanga disponibile un elenco completo delle informazioni a cui fanno riferimento. Si tratta, rispettivamente, dell'origine degli accessi, degli indirizzi richiesti, degli indirizzi referenti, dei programmi usati per accedere, delle stringhe di ricerca e degli utenti (ammesso che l'informazione sia disponibile).

`TopAgents` ***n*** `-A` ***n***	Mostra l'elenco dei programmi usati per accedere, contenente al massimo ***n*** voci.
`TopReferrer` ***n*** `-R` ***n***	Mostra l'elenco degli indirizzi referenti, contenente al massimo ***n*** voci.
`TopSites` ***n*** `-S` ***n***	Mostra l'elenco degli indirizzi di origine, contenente al massimo ***n*** voci.
`TopURLs` ***n*** `-U` ***n***	Mostra l'elenco degli indirizzi richiesti, contenente al massimo ***n*** voci.
`TopCountries` ***n*** `-C` ***n***	Mostra l'elenco delle nazioni di origine, contenente al massimo ***n*** voci.

Direttiva, opzione	Descrizione
`TopEntry` ***n*** `-e` ***n***	Mostra l'elenco delle pagine di ingresso, contenente al massimo ***n*** voci.
`TopExit` ***n*** `-E` ***n***	Mostra l'elenco delle pagine di uscita, contenente al massimo ***n*** voci.
`TopKSites` ***n***	Mostra l'elenco degli indirizzi di origine, in ordine di dimensione dei dati prelevati, contenente al massimo ***n*** voci.
`GroupAgent` ***modello*** ↩ ↪***nome_gruppo***	Dichiara il nome indicato come ultimo argomento della direttiva, al quale si associa tutto il traffico dei programmi che corrispondono al modello. Si osservi che questa direttiva non rimuove le indicazioni dei programmi che vengono raggruppati in questo modo.
`HideAgent` ***modello***	Questa direttiva si usa normalmente dopo una direttiva '**`GroupAgent`**' corrispondente, con lo scopo di non mostrare i nomi dei programmi usati per accedere, che già vengono raggruppati in qualche modo.
`MangleAgents` ***n*** `-M` ***n***	Fa in modo di controllare una prima aggregazione dei programmi usati per accedere. Il numero va da zero a cinque, dove zero richiede di avere tutte le informazioni, mentre cinque le riduce al minimo.

Direttiva, opzione	Descrizione

`CountryGraph no|yes`
 	
`-Y`
	Consente di eliminare il grafico delle nazioni di origine.

Di solito, l'utilizzo di Webalizer è abbastanza semplice, salva l'attenzione che deve essere data al file di configurazione. L'eseguibile che compie il lavoro è '**`webalizer`**', la cui sintassi generale è la seguente:

```
webalizer [opzioni] [file_delle_registrazioni]
```

Alcune delle opzioni sono state descritte a proposito della configurazione; inoltre, come già è stato visto, il file delle registrazioni da analizzare può essere specificato nella configurazione e il file di configurazione può essere indicato espressamente con l'opzione '**`-c`**':

```
-c file_di_configurazione
```

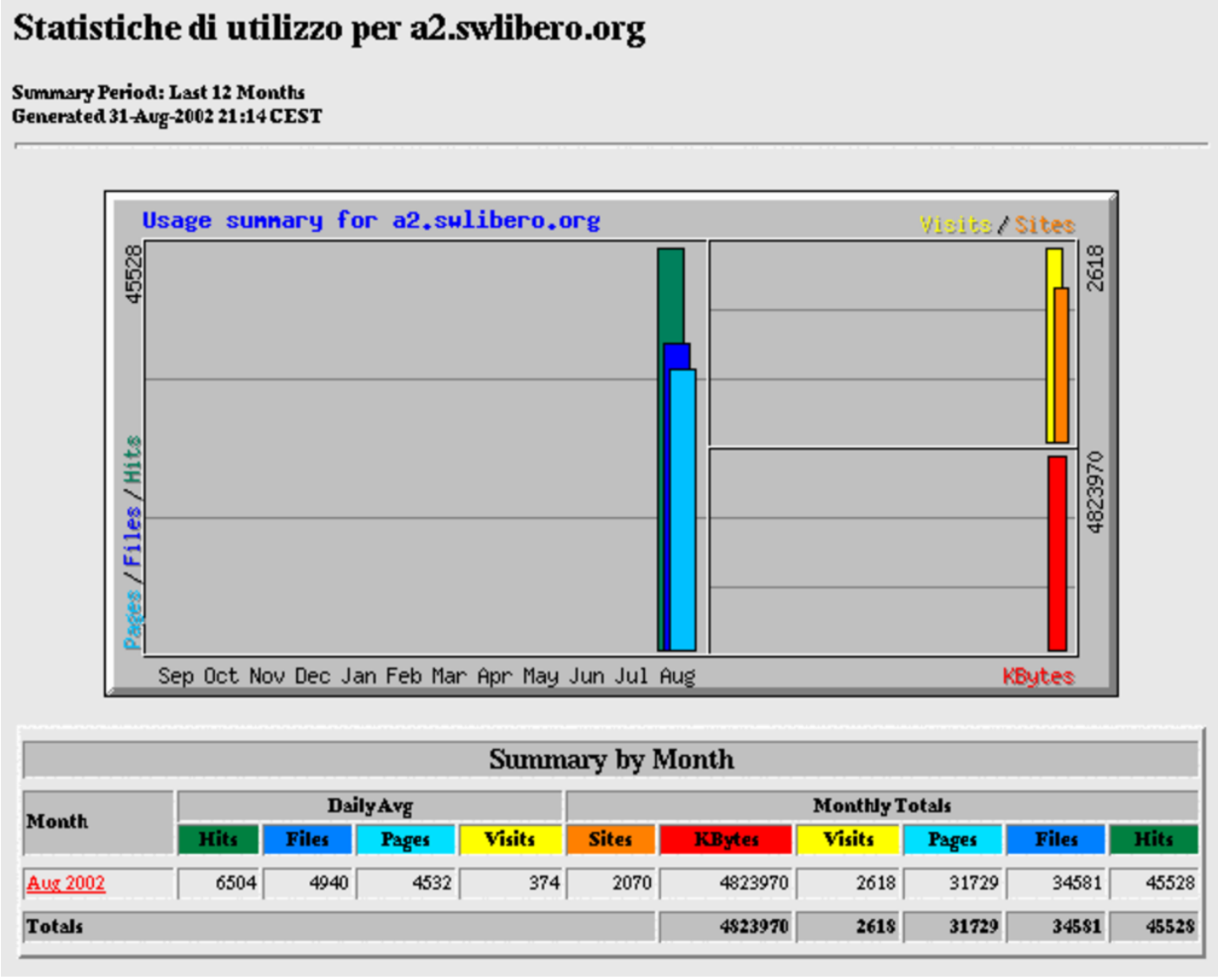
Potendo di indicare il file di configurazione nella riga di comando, è possibile generare statistiche differenti, in base ai contesti di interesse.

In generale, conviene avviare l'eseguibile '**`webalizer`**' specificando sempre il file di configurazione, in modo tale da non dover mettere altro nella riga di comando, curando solo il contenuto della configurazione, come nell'esempio seguente:

```
# webalizer -c /var/www/webalizer.conf [Invio]
```

Naturalmente, in questo modo, nel file di configurazione bisogna stabilire necessariamente la directory in cui devono essere create le statistiche. Le figure seguenti mostrano alcune porzioni di un esempio di statistica generata da Webalizer.

Figura 40.110. La pagina '`index.html`' generata da Webalizer.



Statistiche di utilizzo per a2.swlibero.org

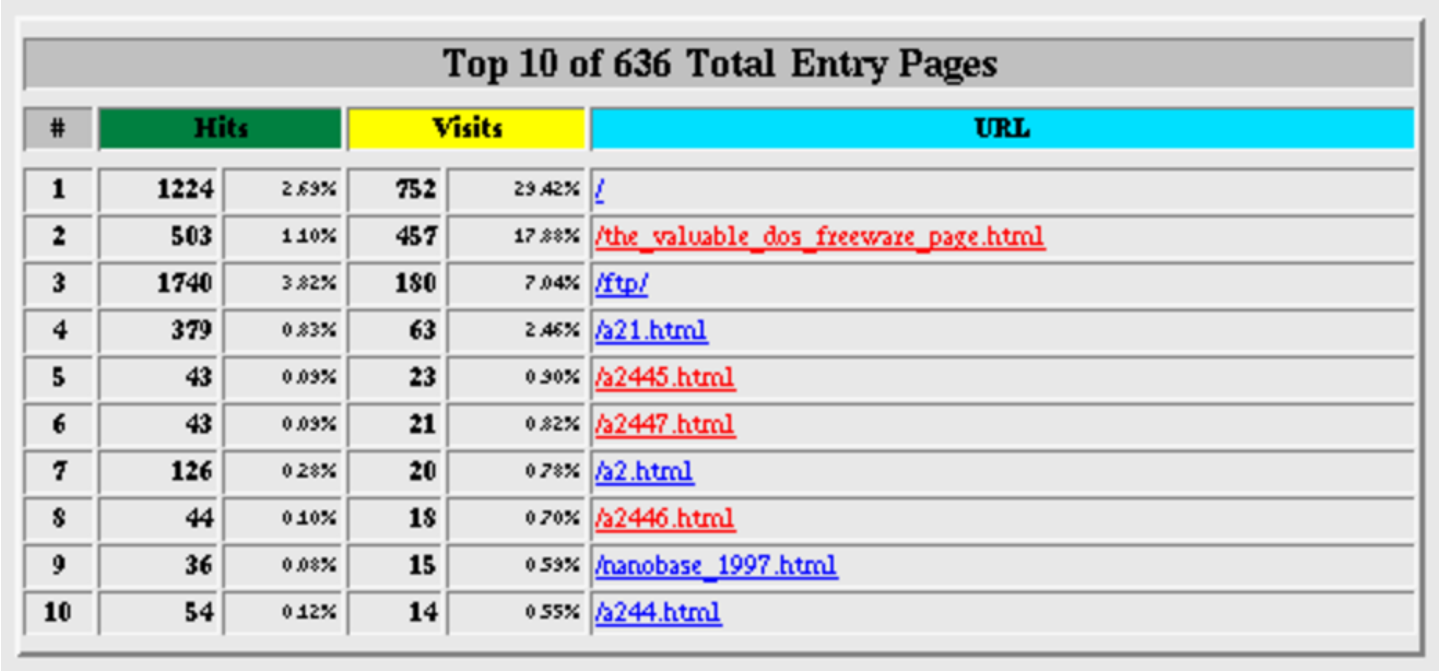
Summary Period: Last 12 Months
Generated 31-Aug-2002 21:14 CEST

Summary by Month

Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
Aug 2002	6504	4940	4532	374	2070	4823970	2618	31729	34581	45528
Totals						**4823970**	**2618**	**31729**	**34581**	**45528**

La pagina iniziale delle statistiche che si ottengono, mostra un riassunto mensile, con una media giornaliera degli accessi. Selezionando il riferimento ipertestuale corrispondente al nome di un mese, se ne ottengono maggiori dettagli.

Figura 40.111. All'interno delle statistiche di un mese, è interessante sapere quali sono le risorse richieste più di frequente.

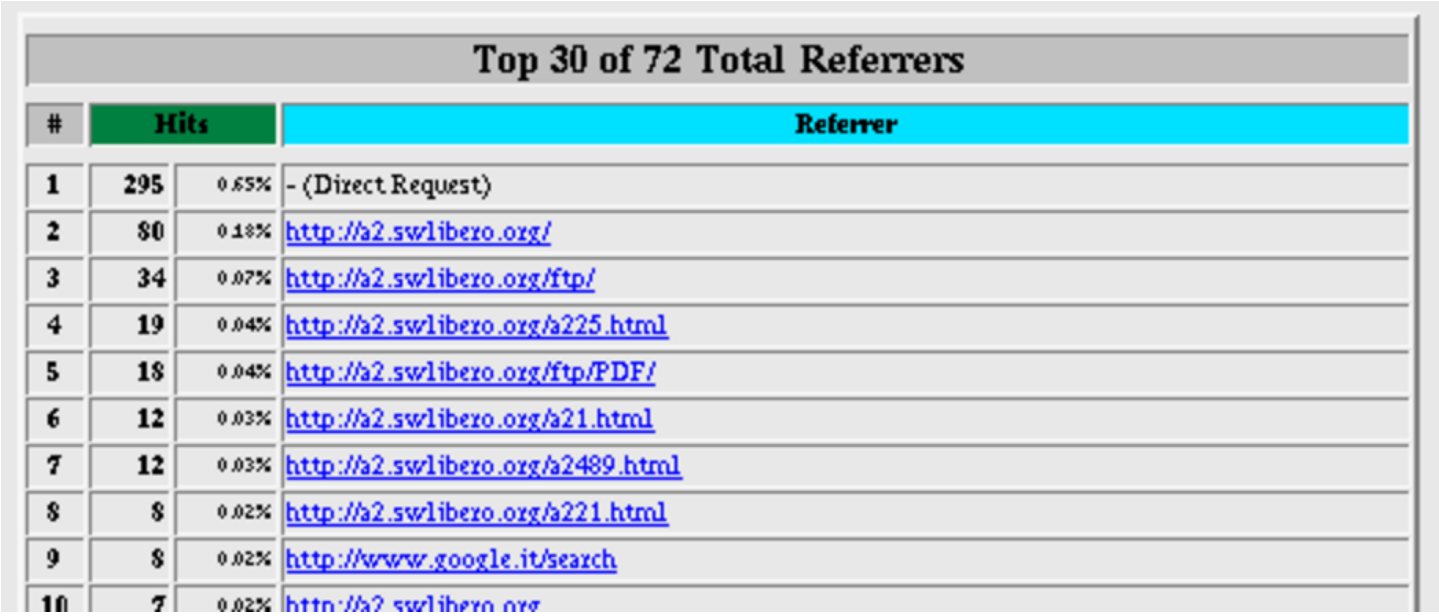


Top 10 of 636 Total Entry Pages

#	Hits		Visits		URL
1	1224	2.69%	752	29.42%	/
2	503	1.10%	457	17.88%	/the_valuable_dos_freeware_page.html
3	1740	3.82%	180	7.04%	/ftp/
4	379	0.83%	63	2.46%	/a21.html
5	43	0.09%	23	0.90%	/a2445.html
6	43	0.09%	21	0.82%	/a2447.html
7	126	0.28%	20	0.78%	/a2.html
8	44	0.10%	18	0.70%	/a2446.html
9	36	0.08%	15	0.59%	/nanobase_1997.html
10	54	0.12%	14	0.55%	/a244.html

La figura precedente mostra in particolare le «pagine di ingresso», o presunte tali. Si tratta in pratica di quelle pagine a cui un utente accede all'inizio della sua visita. Si tratta probabilmente di risorse a cui si arriva attraverso dei segnalibri, oppure dei riferimenti da altri siti.

Figura 40.112. L'elenco dei referenti (si ottiene questa informazione solo se i dati di partenza sono in formato «combinato»).

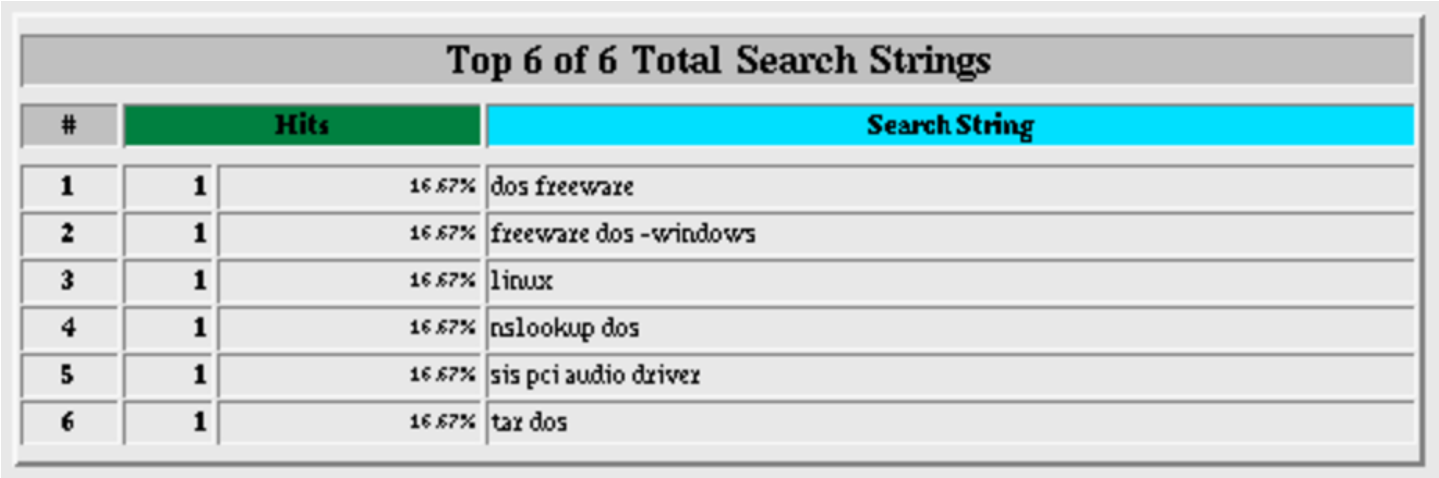


Top 30 of 72 Total Referrers

#	Hits		Referrer
1	295	0.65%	- (Direct Request)
2	80	0.18%	http://a2.swlibero.org/
3	34	0.07%	http://a2.swlibero.org/ftp/
4	19	0.04%	http://a2.swlibero.org/a225.html
5	18	0.04%	http://a2.swlibero.org/ftp/PDF/
6	12	0.03%	http://a2.swlibero.org/a21.html
7	12	0.03%	http://a2.swlibero.org/a2489.html
8	8	0.02%	http://a2.swlibero.org/a221.html
9	8	0.02%	http://www.google.it/search
10	7	0.02%	http://a2.swlibero.org

La figura precedente mostra l'elenco degli indirizzi di provenienza per l'ingresso dei visitatori. In questo caso, trattandosi delle statistiche di *http://a2.swlibero.org*, si manifesta una carenza nella configurazione, dove sarebbe stato meglio mascherare i referenti appartenenti al dominio *a2.swlibero.org*. Comunque, si può vedere nell'esempio che uno dei referenti è un noto motore di ricerca.

Figura 40.113. L'elenco delle stringhe di ricerca estrapolate dagli indirizzi referenti.



Top 6 of 6 Total Search Strings

#	Hits		Search String
1	1	16.67%	dos freeware
2	1	16.67%	freeware dos -windows
3	1	16.67%	linux
4	1	16.67%	nslookup dos
5	1	16.67%	sis pci audio driver
6	1	16.67%	tar dos

I motori di ricerca, quando vengono interpellati, utilizzano solitamente una modalità GET, in modo tale da riportare la stringa di ricerca nello stesso URI contenente l'elenco degli indirizzi che potrebbero corrispondere a ciò che si sta cercando. In tal modo, queste stringhe di ricerca possono apparire come indirizzi referenti; ma se Webalizer riesce a riconoscerle, genera una statistica speciale delle parole o delle stringhe cercate che hanno portato al sito. Nella figura precedente si vede che Webalizer è riuscito a individuare delle stringhe di ricerca dagli indirizzi dei referenti, appartenenti a motori di ricerca noti.

40.8 Wget

«

Il programma Wget[10] è in grado di prelevare file utilizzando sia il protocollo HTTP, sia FTP. La sua caratteristica più importante è la capacità di operare sullo sfondo, senza bisogno di un terminale attivo. In questo senso, è anche insensibile al segnale '**`SIGHUP`**'.[11]

Wget è predisposto normalmente per il prelievo di un file singolo; per questa ragione, in condizioni normali, quando si fa riferimento

a una directory, ammesso che si ottenga l'elenco del suo contenuto, Wget produce un file HTML con tale elenco.

A seconda del fatto che si usi Wget per prelevare materiale attraverso il protocollo HTTP o FTP, il suo comportamento può essere differente; in particolare, quando si utilizza l'FTP, è possibile l'indicazione di metacaratteri (caratteri jolly) per fare riferimento a un gruppo di file.

La scansione ricorsiva deve essere richiesta in modo esplicito attraverso le opzioni o la configurazione, ma mentre nel caso dell'FTP si tratta di un processo abbastanza intuitivo attraverso cui si discendono le varie directory, quando si utilizza il protocollo HTTP ciò significa seguire i riferimenti ipertestuali che si incontrano.

Quando si utilizza Wget per replicare un'area FTP particolare, va tenuto in considerazione il fatto che nella destinazione non vengono eliminati i file che nell'origine invece sono stati rimossi.

40.8.1 Forma dell'URI

Per raggiungere gli oggetti che si vogliono scaricare si utilizzano degli URI, la cui forma può essere espressa dalle sintassi seguenti.

`http://`***nodo*** [`:`***porta***] `/` [***percorso***]

`ftp://`***nodo*** [`:`***porta***] `/` [***percorso***]

```
http://utente [:parola_d'ordine]@nodo [:porta]/[percorso]
```

```
ftp://utente [:parola_d'ordine]@nodo/[percorso]
```

Generalmente, con il protocollo HTTP, l'indicazione di un utente e di una parola d'ordine non è richiesta e di conseguenza si salta. Nel caso del protocollo FTP è invece obbligatoria l'identificazione: quando queste informazioni non vengono fornite, né nell'URI, né nelle opzioni e nemmeno nei file di configurazione, si utilizza il noto utente anonimo ('**ftp**').

Come accennato, l'utente e la parola d'ordine possono essere forniti attraverso opzioni della riga di comando o direttive dei file di configurazione. A questo proposito, è importante osservare che si gestiscono due coppie diverse di nominativo-utente e parola d'ordine: una per il protocollo FTP e una per HTTP.

Bisogna ricordare che l'indicazione della parola d'ordine nella stessa riga di comando (nell'URI o nelle opzioni) è pericolosa perché risulta visibile nell'elenco dei processi in esecuzione.

40.8.2 File di configurazione

«

Wget può essere configurato attraverso due file: '`/etc/wgetrc`' e '`~/.wgetrc`'. Il primo rappresenta la configurazione dell'intero sistema e potrebbe essere collocato anche in un'altra posizione del file system, a seconda della particolare distribuzione GNU che si

utilizza; il secondo è quello personale dell'utente. Le direttive contenute nel file di configurazione personale prevalgono su quelle della configurazione globale di sistema, ma le opzioni della riga di comando prevalgono a loro volta sulla configurazione.

Il contenuto di questi due file di configurazione segue le stesse regole sintattiche. I commenti sono preceduti dal simbolo '**#**' e così sono ignorate anche le righe bianche. Le direttive vengono espresse in forma di assegnamento di variabile, come indicato di seguito:

```
nome = valore
```

Per la precisione si distingue tra direttive che si riferiscono a modalità di funzionamento che possono essere attivate o disattivate, dove si assegnano le parole chiave '**on**' oppure '**off**', da quelle in cui deve essere assegnata una stringa contenente una qualche informazione. In particolare, in questo ultimo caso, se si indica una direttiva in cui non si assegna alcun valore, si intende azzerare implicitamente quanto definito precedentemente per quella funzione di Wget, ma lo stesso ragionamento vale naturalmente anche per le opzioni della riga di comando.

40.8.3 Utilizzo del programma

```
wget [opzioni] uri...
```

Wget si materializza in pratica nell'eseguibile '**wget**'. Come si può vedere dalla sintassi, l'uso di questo programma può essere molto semplice. È necessaria l'indicazione di almeno un URI e in mancan-

za di altri dati si intende ottenere solo la copia dell'oggetto a cui fa riferimento l'URI stesso.

La cosa più importante e delicata che può essere regolata attraverso le opzioni è la scansione ricorsiva del punto di origine, soprattutto quando l'URI di partenza fa riferimento al protocollo HTTP.

L'eseguibile '**wget**' è esente da segnali '**SIGHUP**' e per questo è adatto particolarmente all'uso sullo sfondo (*background*), ma in tal caso è sempre meglio utilizzare '**nohup**' per sicurezza, perché alcune shell provvedono a eliminare i processi loro discendenti quando loro stesse terminano di funzionare.

La sintassi indicata è solo una semplificazione; in realtà, l'URI, pur essendo un'informazione necessaria, potrebbe essere fornito attraverso un file locale contenente uno o più riferimenti da scandire.

La tabella seguente elenca alcune opzioni elementari, assieme alle direttive corrispondenti dei file di configurazione.

Opzione o direttiva	Descrizione
`-o` ***file*** `--output-file=`***file***	Durante il suo funzionamento, vengono generati dei messaggi che normalmente sono emessi attraverso lo standard output. Per evitare che ciò avvenga si può utilizzare questa opzione in modo da creare il file indicato, mettendoci dentro tali messaggi. Se questo file dovesse esistere già, verrebbe cancellato.
`-a` ***file*** `--append-output=`***file***	Invia nel file indicato i messaggi che altrimenti sarebbero destinati allo standard output, come con l'opzione '**-o**', con la differenza che i dati vengono aggiunti al file, se questo esiste già.

Opzione o direttiva	Descrizione

`-v`
`--verbose`
`verbose = on`
	Attiva la modalità dettagliata in cui tutte le informazioni vengono emesse. A meno che il programma sia stato compilato in modo particolare, si tratta sempre della modalità predefinita.

`-nv` `verbose = off`	Questa opzione, permette di disattivare la modalità dettagliata, facendo in modo che siano generati solo i messaggi essenziali.

`-r`
`--recursive`
`recursive = on`
	Questa opzione permette di eseguire una scansione ricorsiva.

`-l` ***n_livelli*** `--level=`***n_livelli*** `reclevel =` ***n_livelli***	Specifica la profondità massima di ricorsione. Questa indicazione è fondamentale quando si vuole riprodurre un URI di tipo HTTP, perché i riferimenti possono andare in ogni direzione. Il valore predefinito è di cinque livelli.

Opzione o direttiva	Descrizione

`-nc`
`--no-clobber`
`noclobber = on`
	In condizioni normali, quando si esegue una scansione ricorsiva allo scopo di prelevare una copia di un URI remoto, i file che dovessero essere già presenti nel sistema locale, verrebbero sovrascritti. Utilizzando questa opzione, si evita la sovrascrittura, ma soprattutto si evita che questi vengano caricati dal nodo remoto. Se si tratta di file HTML, cioè file da cui si può partire per un livello di ricorsione successivo, questi vengono semplicemente letti dal sistema locale.
	In tal modo, questa opzione è importante per riprendere lo scarico di un URI remoto che in precedenza è stato interrotto.

`-t` ***n_tentativi*** `--tries=`***n_tentativi*** `tries =` ***n_tentativi***	Permette di definire un numero di tentativi per accedere alla risorsa. Se si utilizza il numero zero, o la parola chiave '**`inf`**', si intende fare in modo che '**`wget`**' tenti all'infinito.
`-P` ***directory_locale*** `--directory-prefix=`↩ ↪***directory_locale*** `dir_prefix =` ↩ ↪***directory_locale***	Permette di definire una posizione diversa dalla directory corrente per lo scarico dei file dall'URI remoto.

Gli esempi seguenti partono dal presupposto che non sia stato predisposto alcun file di configurazione, per cui tutto quanto è descritto

dalla riga di comando.

- `$` **`wget "http://dinkel.brot.dg/listino.html"`** [*Invio*]

  Preleva il file ‘`listino.html`’ dall’URI ‘`http://dinkel.brot.dg/listino.html`’, salvandolo nella directory corrente.

- `$` **`wget "ftp://dinkel.brot.dg/pub/listino.html"`** [*Invio*]

  Preleva il file ‘`listino.html`’ dall’URI ‘`ftp://dinkel.brot.dg/pub/listino.html`’, salvandolo nella directory corrente.

- `$` **`wget "http://dinkel.brot.dg/"`** [*Invio*]

  Genera il file ‘`index.html`’ nella directory corrente, contenente quanto restituito dall’URI ‘`http://dinkel.brot.dg/`’ (potrebbe trattarsi effettivamente dell’elenco del contenuto oppure di una pagina di ingresso).

- `$` **`wget "ftp://dinkel.brot.dg/"`** [*Invio*]

  Genera il file ‘`index.html`’ nella directory corrente, contenente l’elenco del contenuto dell’URI ‘`ftp://dinkel.brot.dg/`’.

- `$` **`wget -r "ftp://dinkel.brot.dg/pub/progetto/"`** [*Invio*]

  Riproduce l’URI ‘`ftp://dinkel.brot.dg/pub/progetto/`’ con tutto il contenuto della directory specificata e di quelle successive fino al massimo numero di livelli predefinito (cinque), generando il percorso ‘`./dinkel.brot.dg/pub/progetto/...`’ nella directory corrente.

- `$` **`wget -r -l inf "ftp://dinkel.brot.dg/pub/progetto/"`** [*Invio*]

  Come nell’esempio precedente, ma viene riprodotto tutto il ramo ‘`progetto/`’, senza limiti di livelli di ricorsione. Infatti, trattan-

dosi del protocollo FTP, non si pongono problemi a questo tipo di scelta, dal momento che la struttura ha un termine.

- `$ wget -r -l inf -nc ↵`
  `↪ "ftp://dinkel.brot.dg/pub/progetto/"` [*Invio*]

  Come nell'esempio precedente, con la differenza che, se parte dei file contenuti nell'URI remoto sono già presenti localmente, questi non vengono prelevati effettivamente.

- `$ nohup wget -r -l inf -nc -o ~/mio_log ↵`
  `↪ "ftp://dinkel.brot.dg/pub/progetto/" &` [*Invio*]

  Come nell'esempio precedente, con la differenza che il processo viene messo sullo sfondo (*background*) e viene controllato da '**nohup**', in modo da garantire che non sia interrotto quando la shell termina di funzionare. Inoltre viene generato il file '`~/mio_log`' con i messaggi emessi.

- `$ wget -r "http://dinkel.brot.dg/progetto/"` [*Invio*]

  Riproduce l'URI '`http://dinkel.brot.dg/progetto/`' con tutto il contenuto, in base ai riferimenti che vengono incontrati, fino al massimo numero di livelli predefinito (cinque), generando il percorso '`./dinkel.brot.dg/progetto/...`' nella directory corrente.

- `$ wget -r -nc "http://dinkel.brot.dg/progetto/"` [*Invio*]

  Come nell'esempio precedente, ma i file già esistenti non vengono prelevati nuovamente e di conseguenza non vengono sovrascritti.

40.8.4 Scansione a partire da un file locale

L'eseguibile '**wget**' permette di non indicare alcun URI nella riga di comando, utilizzando al suo posto l'inclusione di un file locale. Questa modalità viene utilizzata normalmente in modo congiunto a quella ricorsiva, ottenendo la scansione di tutti gli indirizzi URI contenuti nel file.

Il file può essere in formato HTML (è la cosa migliore) e in tal caso vengono seguiti i riferimenti ipertestuali, altrimenti può andare bene anche un file di testo contenente un elenco di indirizzi puri e semplici. Il problema si pone semmai quando il file indicato è in HTML, ma incompleto; in questo caso occorre specificare con un'opzione apposita che deve essere interpretato come HTML.

Gli indirizzi URI dovrebbero essere assoluti; se non lo sono, si può utilizzare un'opzione apposita per indicare l'URI di partenza, oppure, se si tratta di un file HTML, si può aggiungere un elemento speciale:

```
<base href="uri">
```

Tuttavia, è bene tenere presente che si tratta di un elemento non previsto nel DTD dell'HTML, quindi va usato solo in questa circostanza.

Opzione o direttiva	Descrizione
`-i` ***file*** `--input-file=`***file*** `input =` ***file***	Permette di indicare il file (HTML o un semplice elenco di URI) da utilizzare come punto di partenza per una scansione ricorsiva.

Opzione o direttiva	Descrizione

`-F`
`--force-html`
`force_html = on`
	Richiede di interpretare il file indicato come HTML.

`--base=`***uri*** `base =` ***uri***	Specifica esplicitamente un URI di partenza per i riferimenti relativi contenuti nel file.

Segue la descrizione di alcuni esempi.

- `$` **`wget -r -i elenco.html`**[*Invio*]

  Scandisce tutti i riferimenti che trova nel file '`elenco.html`'.

- `$` **`wget -r -i elenco --force-html`**[*Invio*]

  Come nell'esempio precedente, con la differenza che il file '`elenco`' non viene riconosciuto automaticamente come HTML, per cui è stata aggiunta l'opzione '**`--force-html`**'.

- `$` **`wget -r -i elenco --base="http://dinkel.brot.dg/"`**[*Invio*]

  Viene scandito il file '`elenco`' (il tipo di questo viene determinato in modo automatico), ma in più viene specificato che gli indirizzi relativi hanno il prefisso '`http://dinkel.brot.dg/`'.

40.8.5 Scansione ricorsiva

La scansione ricorsiva di un URI è ciò che genera i problemi maggiori nella gestione di Wget, cosa che dovrebbe essere già stata compresa dall'esposizione fatta fino a questo punto. La scansione ricorsiva di un URI di tipo FTP è abbastanza intuitiva, dal momento che si riferisce a un ramo di directory, mentre quando si tratta di un URI di tipo HTTP, questa ricorsione si basa sui riferimenti '**`HREF`**' e '**`SRC`**'; quando poi il file scaricato è di tipo '**`text/html`**', questo viene scandito alla ricerca di altri riferimenti da seguire.

Soprattutto quando si opera con il protocollo HTTP, è importante porre un limite alla ricorsione, dal momento che i riferimenti possono articolarsi in modi imprevedibili. Ma oltre a questo, può essere conveniente limitare la scansione ricorsiva ai riferimenti relativi, oppure a quelli di un dominio particolare.

Quando la scansione ricorsiva è normale, cioè non si limita ai soli riferimenti relativi, si pone il problema di trattare convenientemente i riferimenti ipertestuali assoluti che puntano allo stesso nodo in cui si trovano. Infatti, può accadere che due nomi si riferiscano allo stesso nodo; in tal caso non ha senso sdoppiare i percorsi, anche perché si rischierebbe di duplicare lo scarico di alcuni file. Per risolvere questo problema, Wget interpella il sistema DNS in modo da verificare se si tratta della stessa macchina o meno.

La vera difficoltà nasce quando il servente HTTP distingue tra nodi virtuali differenti, a cui corrisponde però lo stesso indirizzo IP, in base all'uso di un diverso alias per raggiungere lo stesso elaboratore. In tal caso, occorre informare Wget di ignorare il sistema DNS e limitarsi al confronto letterale dei nomi dei nodi.

Opzione o direttiva	Descrizione

`-L`
 	
`--relative`
 	
`relative_only = on`
	Fa in modo di seguire solo i riferimenti relativi, escludendo quindi qualunque URI completo dell'indicazione del nodo.

`-np`
 	
`--no-parent`
 	
`no_parent = on`
	Permette di evitare che siano attraversate directory precedenti a quella dell'URI di partenza.

`-X` ***elenco_directory*** `--exclude` ***elenco_directory*** `exclude_directories = ` ↩ ↪***elenco_directory***	Permette di escludere un elenco di directory dalla scansione ricorsiva.

Opzione o direttiva	Descrizione

`-nH`
 	
`add_hostdir = off`
	Disabilita la creazione di directory locali prefissate dal nome del nodo di origine. Di solito, in presenza di una scansione ricorsiva di un URI, viene creata localmente una struttura di directory che riproduce il sistema remoto, a partire dal nome del nodo stesso.
	Questa opzione è utile solo quando si è sicuri che i riferimenti non si sviluppano all'indietro (eventualmente attraverso l'uso di opzioni opportune), come quando si opera con URI di tipo FTP.

`-nh`	Disabilita il controllo DNS; in tal modo non viene verificato se due nomi a dominio appartengono in realtà allo stesso nodo.
`-D` ***elenco_domini*** `--domains=`***elenco_domini*** `domains =` ***elenco_domini***	Permette di definire un elenco di domini accettabili. In pratica, si permette a Wget di seguire i riferimenti a nodi differenti da quello di partenza, purché appartengano ai domini elencati.

`-k`
 	
`--convert-links`
 	
`convert_links = on`
	In questo modo si ottiene di convertire i riferimenti assoluti in riferimenti relativi, limitatamente ai file scaricati effettivamente.

Segue la descrizione di alcuni esempi.

- `$` **`wget -r -L -np "http://dinkel.brot.dg/progetto/"`** [*Invio*]

  Riproduce l'URI '`http://dinkel.brot.dg/progetto/`' con tutto il contenuto, in base ai riferimenti **relativi** che vengono incontrati, escludendo quelli che si riferiscono a posizioni precedenti alla directory '`/progetto/`', fino al massimo numero di livelli predefinito (cinque), generando il percorso '`./dinkel.brot.dg/progetto/...`' nella directory corrente.

- `$` **`wget -r -L -np "http://dinkel.brot.dg/progetto/" ↩`**
  `↪ -X /progetto/img/,/progetto/x/` [*Invio*]

  Come nell'esempio precedente, con l'aggiunta che non vengono riprodotte le directory '`/progetto/img/`' e '`/progetto/x/`'.

- `$` **`wget -r -D .brot.dg "http://dinkel.brot.dg/"`** [*Invio*]

  Riproduce l'URI '`http://dinkel.brot.dg/progetto/`' seguendo anche i riferimenti ad alti nodi purché appartenenti al dominio *.brot.dg*.

40.8.6 Selezione dei file in base al loro nome

«

Quando si scandisce un URI remoto in modo ricorsivo, è possibile definire i file da scaricare in base al nome. Nel caso particolare del protocollo FTP, si possono utilizzare i noti metacaratteri (caratteri jolly) nello stesso URI, mentre con il protocollo HTTP le cose cambiano perché ci si deve sempre affidare alla scansione dei riferimenti contenuti nelle pagine HTML.

Opzione o direttiva	Descrizione
`-A` ***elenco_da_accettare*** `--accept` ***elenco_da_accettare*** `accept =` ***elenco_da_accettare***	In questo modo si può specificare un elenco di suffissi o di modelli espressi attraverso metacaratteri riferiti a file che si vogliono scaricare. In pratica, si scaricano solo questi file, o meglio, gli altri che sono serviti per raggiungerli vengono rimossi successivamente.
`-R` ***elenco_da_escludere*** `--reject` ***elenco_da_escludere*** `reject =` ***elenco_da_escludere***	In questo modo si può specificare un elenco di suffissi o di modelli espressi attraverso metacaratteri riferiti a file che **non** si vogliono scaricare. Tutti gli altri file vanno bene.

Segue la descrizione di alcuni esempi.

- `$` **`wget -r -A "*.gif,*.jpg" "http://dinkel.brot.dg/progetto/"`** [*Invio*]

  Salva localmente solo i file che terminano per '`.gif`' e '`.jpg`', provenienti dall'URI '`http://dinkel.brot.dg/progetto/`'.

- `$` **`wget -r -R "*.gif,*.jpg"`** ↩
  ↪ **`"http://dinkel.brot.dg/progetto/"`** [*Invio*]

  Come nell'esempio precedente, con la differenza che viene scaricato tutto fuorché i file che terminano per '`.gif`' e '`.jpg`'.

40.8.7 Identificazioni e parole d'ordine

Si è già accennato al fatto che il nome dell'utente e la parola d'ordine eventualmente necessari per accedere a determinati servizi FTP e HTTP possono essere inseriti nello stesso URI. In alternativa si possono usare delle opzioni apposite o delle direttive dei file di configurazione.

È bene ricordare che solo inserendo le parole d'ordine all'interno del file di configurazione personale si può evitare che queste siano visibili, perché se si immettono direttamente nella riga di comando, queste diventano accessibili dall'elenco dei processi, per tutti gli utenti.

Opzione o direttiva	Descrizione
`--http-user` ***utente*** `http_user =` ***utente***	Permette di definire il nominativo-utente da usare per una connessione HTTP a un particolare URI che richiede l'identificazione.
`--http-passwd` ***parola_d'ordine*** `http_passwd =` ***parola_d'ordine***	Permette di definire la parola d'ordine da usare per una connessione HTTP a un particolare URI che richiede l'identificazione.
`passwd =` ***parola_d'ordine***	Permette di definire la parola d'ordine da usare per una connessione FTP.

40.8.8 Riproduzione speculare e informazioni data-orario

Quando si vuole riprodurre un URI remoto e si vuole mantenere la copia locale allineata con quella remota, la cosa più importante da verificare è la variazione dell'informazione data-orario degli oggetti remoti. In pratica, si vuole ottenere che:

- vengano scaricati i file remoti se non sono già presenti nel sistema locale, o se la dimensione non combacia;
- vengano scaricati i file remoti se la loro data di modifica è più recente rispetto a quella dei file locali.

Opzione o direttiva	Descrizione

`-N`
`--timestamping`
`timestamping = on`
	Si fa in modo che venga attuato il meccanismo di aggiornamento in base alla verifica delle date, evitando così di ripetere ogni volta il prelievo di dati già esistenti localmente e presumibilmente aggiornati.

`-m`
`--mirror`
`mirror = on`
	Equivale alla richiesta di una ricorsione infinita assieme all'attivazione di '**`timestamping`**' e '**`noclobber`**'.

L'esempio seguente serve a riprodurre nella directory corrente ciò che si dirama a partire da '`http://dinkel.brot.dg/articoli/`' senza seguire riferimenti in altri nodi, né all'interno di percorsi che si articolano da posizioni precedenti gerarchicamente. In particolare vengono trasformati i riferimenti in modo che siano solo relativi (senza l'indicazione del nodo)

```
# wget --mirror --relative --no-parent ↩
↪      -nH "http://dinkel.brot.dg/articoli/" [Invio]
```

Questo esempio rappresenta l'utilizzo di Wget per ottenere la riproduzione speculare di un'area HTTP. Tuttavia, il difetto di questo approccio sta nel fatto che Wget non è in grado di verificare la scomparsa di file dall'origine, per cui non può provvedere da solo alla loro eliminazione.

40.8.9 Funzionalità varie

«

Altre funzionalità di Wget possono essere molto utili e queste sezioni non esauriscono la descrizione delle possibilità che ci sarebbero. Per approfondire lo studio di Wget occorre consultare la sua documentazione, che normalmente è disponibile in forma di ipertesto Info: *info wget*. La tabella successiva riporta altre opzioni di una certa importanza che non hanno trovato posto nelle altre tabelle analoghe.

Opzione o direttiva	Descrizione
`-c` `--continue`	Permette di riprendere il prelievo di un file (uno solo) continuando da dove l'operazione è stata interrotta precedentemente. Questa opzione è efficace solo se il servente relativo è predisposto per questa funzionalità.

Opzione o direttiva	Descrizione
`-Q` ***dimensione*** `--quota` ***dimensione*** `quota =` ***dimensione***	Permette di definire il limite massimo di spazio utilizzabile per i prelievi, quando questi sono fatti in modo **ricorsivo**. Il valore della dimensione viene espresso da un numero che rappresenta una quantità di byte. Se questo numero è seguito dalla lettera '**`k`**', indica unità in kibibyte (simbolo: «Kibyte»), altrimenti, se è seguito dalla lettera '**`m`**', si riferisce a unità in mebibyte (simbolo: «Mibyte»).
`--spider`	In questo modo si ottiene soltanto la verifica che l'URI indicato rappresenti un oggetto esistente. Se l'oggetto non esiste, o non è raggiungibile, l'eseguibile '**`wget`**' termina di funzionare restituendo un valore diverso da zero, cosa che può servire per costruire degli script per la verifica di un elenco di URI, per esempio quello di un segnalibro di un programma di navigazione. Purtroppo, tale funzionalità non si adatta bene al protocollo FTP.
`-w` ***n_secondi*** `--wait` ***n_secondi*** `wait =` ***n_secondi***	Permette di stabilire un intervallo di tempo tra il prelievo di un file e il successivo. È molto utile per alleggerire il carico del sistema locale, di quello remoto e dell'utilizzo della banda.

40.9 Riferimenti

«

- *W3C, World Wide Web Consortium*, *http://www.w3.org/*
- Michiel Boland, *Mathopd*, *http://www.mathopd.org/*
- Ian Graham, *Web/HTML Documentation and Developer's Resource*, *http://www.utoronto.ca/webdocs/*
- Christian Neuss, Johan Vromans, *Perl, guida pratica*, Apogeo, 1996
- Steven Brenner, '**`cgi-lib.pl`**', libreria standard per la creazione di script CGI in Perl, *http://cgi-lib.berkeley.edu/*
- *ht://Dig*, *http://www.htdig.org/*
- *Webalizer*, *http://www.mrunix.net/webalizer/*

[1] **W3M** software libero con licenza speciale

[2] **Mathopd** software libero con licenza speciale

[3] L'uso del punto interrogativo rende la cosa intuitiva: la richiesta viene fatta attraverso un'interrogazione.

[4] I motori di ricerca utilizzano normalmente il metodo '**`GET`**', perché consente di trasmettere l'interrogazione richiesta nell'indirizzo usato, il quale viene memorizzato dai serventi HTTP come referente. Questa è una situazione pratica in cui il metodo '**`POST`**' non sarebbe adatto.

[5] L'esempio del file '`form-test.html`' viene proposto secondo lo standard HTML 4.01, perché alcuni attributi usati sono incompatibili con ISO-HTML.

[6] **ht://Dig** GNU GPL

[7] La dichiarazione del modulo, con l'elemento '**`FORM`**' va verificata per quanto riguarda l'attributo '**`ACTION`**', che deve puntare esattamente al programma CGI di ht://Dig, presso il sito che interessa.

[8] Eventualmente, le statistiche di accesso possono servire anche per dimostrare la visibilità reale di pagine a contenuto pubblicitario, ma rimane il fatto che sia facile creare dei file di registrazioni fasulli per ingannare i finanziatori.

[9] **Webalizer** GNU GPL con l'uso di una libreria che ha una licenza differente

[10] **Wget** GNU GPL

[11] Alcune shell, quando concludono la loro attività, cercano di eliminare i processi loro discendenti, senza limitarsi a inviare un semplice '**`SIGHUP`**'. In tal caso conviene avviare '**`wget`**' attraverso '**`nohup`**'.