Tema 4 - Client HTTP

Scrieti in C sau C++ un program client care se conecteaza la un server HTTP si downloadeaza pagina web ceruta de utilizator, existand posibilitatea de download recursiv (vezi specificatiile de mai jos). Un program cu functionalitate asemanatoare, dar mai extinsa, este *wget* (*man wget* pentru detalii).

Specificatii:
Programul se va apela din linia de comanda astfel:

./myclient [-r] [-e] [-o <fisier_log>] http://<nume_server>/<cale_catre_pagina>
Optiunile au urmatoarele semnificatii:

- *-r (recursive)*: daca aceasta optiune este activata, programul va realiza download recursiv, adica va parcurge pagina web si in cazul in care intalneste link-uri, downloadeaza si paginile referite de acestea. Se presupune ca link-urile se refera numai la pagini de pe acelasi server, deci nu vor avea forma *...*, ci vor fi doar de forma *...*. Fisierele referite de linkuri vor fi doar in format html. Pe calculatorul clientului se va crea o structura de directoare si fisiere asemanatoare cu cea de pe server, directorul parinte al acesteia avand nume identic cu al serverului. De exemplu, daca se cere downloadarea paginii *http://site/test/dir1/ceva.html*, pe calculatorul clientului pagina va avea urmatoarea cale: *site/test/dir1/ceva.html* (iar directorul *site* se va afla in directorul curent)
- -e (everything): daca aceasta optiune este activata se vor downloada toate fisierele la care se face referire in paginile html printr-un link de forma .... Se considera ca numele fisierelor vor avea o extensie de 3 sau 4 caractere. Aceasta optiune se poate utiliza in combinatie cu optiunea -r si se vor downloada toate fisierele .zip, .pdf, etc. pentru care exista link-uri in paginile parcurse.
- *-o <fisier_log>*: aceasta optiune specifica scrierea mesajelor de eroare intr-un fisier de log. Daca optiunea nu este activata, mesajele de eroare se scriu la iesirea standard de eroare (stderr). Atentie! Trebuie sa generati mesaje pentru toate cazurile in care apar erori, si sa specificati in aceste mesajele cauzele erorilor. Daca scrieti programul in C++ este de preferat sa utilizati exceptii.

Precizari:

Nivelul maxim de recursivitate pe care trebuie sa il suporte clientul este 5 (se considera ca pagina initiala este pe nivelul 1, paginile referite din ea pe nivelul 2 etc.)

Veti downloada recursiv numai link-urile catre fisiere html (extensia .html sau .htm) de pe acelasi server, care pot avea una din urmatoarele forme:

...

...

...

Link-urile catre pagini de pe alte servere (care incep cu "http://") sau catre sectiuni de pagina (care contin caracterul "#") vor fi ignorate. Nu trebuie sa tratati cazul in care pagina HTML contine comentarii (adica atunci cand intalniti un link, nu trebuie sa verificati daca se afla in interiorul unui comentariu).

Pentru a crea directoare din cadrul unui program C sub Linux puteti folosi:

- functia *system()*, care are ca argument o comanda a sistemului de operare (in acest caz va fi comanda *mkdir*)
- apelul de sistem *mkdir* (*man 2 mkdir*)