# Laboratorio Online Permanente di Tecnologie Internet per la Scuola – #loptis

## Un cMOOL: connectivist Massive Open Online Laboratory

OCTOBER 20, 2014

# Copiare pagine Web nella PirateBox – #loptis Reply

*#loptis*, *PirateBox* • *Tags: HTTrack*, *PirateBox*, *salvare*, *save*, *siti web*, *wget*, *WinHTTrack*
L'operazione sembrerebbe a prima vista semplice: prima salvo sul mio computer la pagina web che m'interessa e poi la carico sulla penna USB della PirateBox, nella cartella opportuna. Ma, "salvare una pagina web", non è come salvare un documento tipo Office (LibreOffice, OpenOffice, Microsoft ecc.), a meno che non si tratti di semplici pagine HTML, cosa abbastanza rara oggi.

Molti sanno che i programmi di navigazione hanno un comando di salvataggio delle pagine Web. In Firefox per esempio si trova nel menu **File** (il primo a destra in alto), e poi **Salva pagina con nome…**. D'altro canto, in realtà i siti più visitati sono quasi sempre *servizi web*, qualcosa di molto più complicato di una pagina HTML. Sono sistemi che consentono di elaborare le informazioni tratte da repositori sterminati: consultazione di orari ferroviari, prenotazione di viaggi, ricerca di risorse bibliografiche, servizi didattici di calcolo scientifico, giochi online di massa, social network, e via dicendo. Che cosa combina allora quel semplice comando, **Salva pagina con nome…**?

Un servizio web è composto da uno o più database, una moltitudine di file scritti in HTML, parti di software che girano sul computer che lo ospita (server) e parti di software che vengono caricate sul mio computer (client) insieme alle pagine stesse (spiegazione divulgativa in Due o tre cose sulle pagine Web (http://iamarf.org/2013/12/12/due-o-tre-cose-sulle-pagine-web-loptis/)). Quando clicco qualcosa, il software sul server esegue varie operazioni e, in maniera dipendente dalle circostanze prima compone la pagina che ho richiesto e poi me la invia, non già perfettamente confezionata ma in una forma ancora non finita, diciamo una sorta di kit, come un pacco Ikea, che contiene i pezzi dell'oggetto e gli attrezzi per farlo funzionare. Solo che qui la faccenda è un po' più complicata, e il mio computer, una volta diligentemente montata la pagina, continua a comunicare con il server per rendere la pagina bella dinamica e interattiva ai miei occhi. Insomma, una gran lavorio, anche solo per cercare l'orario di un treno.

Supponiamo ora di avere visualizzato http://iamarf.org (http://iamarf.org) e di applicare il comando **Salva pagina con nome…** – così facendo è sicuro che salverò un file sul mio disco rigido, e sarà un file

HTML, che potrò caricare nel navigatore con il comando File -> Apri file… certamente. Ma non potrò certo pensare che l'operazione di caricamento e visualizzazione di un file, come fosse un documento qualsiasi, possa dar vita a tutto quel processo!

La questione è quindi complessa. Per cercare di rispondere adeguatamente occorre specificare meglio la domanda, distinguendo almeno quattro casi.

## 1. Salvare l'intero sito, con tutte le sue funzionalità. Esempio: ho un blog in WordPress e lo voglio "portare" nella PirateBox, o altrove, così com'è, con tutte le sue funzionalità – scrivere post, commenti ecc.

### L'essenziale

La domanda è legittima, perché la PirateBox è un web server, seppur di dimensioni minime, ma la risposta è no, non si può fare. La PirateBox fatta con il router MR3020 è troppo piccola per sostenere l'elaborazione richiesta da un blog. Se non ti interessano i dettagli puoi saltare al punto successivo.

### L'approfondimento

Capita per esempio che qualcuno chieda: posso trasferire il mio blog nella PirateBox? In primo luogo è chiaro che, anche qualora riuscissi a riprodurre *in toto* la funzionalità del blog, il trasferimento rappresenterebbe comunque una biforcazione, dove nella versione clonata non potrebbero apparire i commenti ricevuti dal blog successivamente alla copia. Allo stesso modo, i commenti fatti alla versione clonata non potrebbero apparire su quella originale.

Ma la difficoltà sostanziale è costituita dalle risorse di cui un blog ha bisogno per funzionare. Il software preferibile per realizzare un blog è WordPress.org (http://wordpress.org), sia perché è il più diffuso al mondo, sia perché ha il pregio di essere distribuito con licenza GPLv2 (o successiva) (http://www.wordpress.org/about/license/), quindi si tratta di software libero. Tali caratteristiche lo rendono utilizzabile in una grande quantità di contesti, ma il caso della nostra PirateBox, basata sul router MR3020, è veramente estremo. In principio, i componenti necessari per far funzionare WordPress potrebbero essere installati nel sistema operativo OpenWRT, versione Linux "embedded" che gira nel router – ovvero un'applicazione web server, un database MySQL e il software PHP. Per il web server la scelta standard è Apache (https://it.wikipedia.org/wiki/Apache_HTTP_Server), l'applicazione di gran lunga più usata – il 70% dei nostri clic viene servito da un server Apache. Ma nel caso di contesti dalle prestazioni limitate è preferibile un'applicazione pensata per essere più leggera: è documentato anche l'uso di WordPress con Lighttpd (https://it.wikipedia.org/wiki/Lighttpd), che è il servizio che gira nelle PirateBox. Il PHP si può installare, anche questo è documentato dal Team PirateBox (http://piratebox.cc/openwrt:mods), anche se in forma sperimentale. Probabilmente anche MySQL, ma di questo sono meno sicuro. In ogni caso, quandanche si riuscisse a far funzionare il tutto, la povera scatolina si inginocchierebbe sotto il peso delle elaborazioni che un tale sistema richiederebbe.

Il discorso potrebbe essere diverso utilizzando una scheda Raspberry Pi (https://it.wikipedia.org/wiki/Raspberry_Pi). C'è molta gente che ha messo su piccoli server con questa scheda. La cosa è certamente possibile, si tratta poi di vedere il carico di lavoro che deve sopportare – certamente non potrà essere eccessivo, ma probabilmente, nell'ambito di una classe o comunque piccolo assembramento, potrebbe essere sufficiente. Da vedere. Da vedere anche se il gioco vale la candela: c'è da lavorarci un po', dipende dalla domanda.

## 2. Non pretendo che funzioni ma che si presenti completo, lasciandomi scendere in tutti i link che mi offre, mi lasci vedere i video, ascoltare i file audio e mi mostri correttamente grafica e immagini

### L'essenziale

Si può fare in parte. Puoi usare l'applicazione WinHTTrack (http://www.httrack.com/) per scaricare una discreta "fotografia" del sito, potendo entrare nei link fino al livello che decidi tu. Ma, tale copia del sito, che WinHTTrack riunisce all'interno di una cartella, è fatto per essere *caricata* dal tuo disco rigido, non per essere *servita* da un server Web, quale la PirateBox è. Tu la puoi offrire attraverso la PirateBox, ma così: *zippi* tutta la cartella prodotta da WinTTRack in un file che poi carichi nella penna USB della PirateBox – i tuoi allievi potranno scaricarla, aprirla sui propri device, ove possibile, e poi navigare il tutto localmente. Ecco le istruzioni passo passo dove supponiamo di scaricare iamarf.org. In verde ciò che devi decidere tu, in rosso ciò che è obbligatorio per ottenere il risultato.

- lanciare WinHTTrack
- **Next**
- **Project name**: **iamarf** (nome a piacere)
- **Directory: My Web Sites** leggi quella che ti propone oppure sceglila tu, in ogni caso prendi nota di dove è collocata, per ritrovare successivamente il progetto
- **Next**
- **Add URL…**
- **URL: http://iamarf.org**
- **OK**
- **Set options…**
- **Limits**
  - **Maximum mirroring depth: 2**
  - **Maximum external depth: 1**
- **Build**
  - **Local Structure Type (how link are saved): Html in web/, images/other in web/** – occhio a scegliere esattamente questa fra le varie simili
  - **DOS Names (8.3) spunta la casella**
- **Next**
- **Finish**
- Puoi andare a fare qualcos'altro, ci vorrà un po' di tempo – qualche dato nell'approfondimento successivo.

- Quando è finito, all'interno della cartella **My Web Sites** (o quello che è) troverai il tuo sito in una cartella di nome **iamarf.org**. Dentro a questa, con il navigatore, puoi aprire il file **index.html** e percorrere il sito off-line fino alla profondità che avrai scelto (vedi l'approfondimento sotto).
- Per offrire il sito così scaricato attraverso la PirateBox, vai alla fine dell'approfondimento seguente.

## L'approfondimento

I software più noti per scaricare un sito Web, ambedue liberi, sono Wget (https://www.gnu.org/software/wget/) e HTTrack (http://www.httrack.com/). Il primo è un componente standard e "antico" presente in quasi tutte le distribuzioni Linux. Si può usare solo da riga di comando. È quello che uso per fare automaticamente un backup settimanale completo di questo blog. Ma qui ci concentriamo su HTTrack, perché la versione per Windows, WinHTTrack è dotata di interfaccia grafica, familiare ai più.

Chi dopo avere scaricato WinHTTrack, fruga fra le opzioni disponibili, si accorge subito del fatto che scaricare un sito non è uno scherzo. Tant'è che per partorire questo pur smilzo tutorial è stato necessario provare una trentina di combinazioni. Tutto questo tenendo sempre conto del fatto che esplorare offline un sito scaricato in questo modo, **non è** come esplorare lo stesso sito *servito* dal web server, per i motivi che ho cercato di descrivere all'inizio.

Sarebbe dispersivo discutere tutte le opzioni possibili, ma merita soffermarsi un attimo sul concetto di profondità. Con questa si intende fino a che punto si desidera scendere nei link: cliccare su un qualsiasi link della pagina vuol dire scendere di un livello, cliccare un link sulla pagina così raggiunta, significa scendere di due livelli, e così via. In WinHTTrack la profondità è determinata da due parametri: **Maximum mirroring depth** e **Maximum external depth**. Il primo rappresenta la profondità che si vuole raggiungere all'interno del sito e il secondo la profondità che si vuole raggiungere nelle pagine esterne. Sono importanti perché hanno un effetto enorme sul comportamento del programma. La quantità di pagine che vengono scaricate cresce infatti esponenzialmente con la profondità, e basta aumentare di poco per rischiare di scaricare quasi tutta l'internet.

Scaricando per esempio http://iamarf.org (http://iamarf.org) [1]:

- con **Maximum mirroring depth=2** e **Maximum external depth=1** si scaricano **34 MB** in **5 minuti**
- con **Maximum mirroring depth=3** e **Maximum external depth=1** si scaricano **414 MB** in **2 ore e 40 minuti**

Copiare una quantità eccessiva di un sito presenta vari svantaggi: si può facilmente occupare una quantità esagerata del proprio disco rigido, i tempi possono essere facilmente inaccettabili e si può procurare un carico eccessivo al sito che si copia. Consapevoli di tali rischi, alcuni amministratori di Web server programmano i propri sistemi in maniera da evitare tout court di essere copiati con tali sistemi, oppure, se si accorgono che qualcuno impone al server un carico eccessivo lo bloccano mettendolo in una black list. Quindi prudenza!

## Per offrire il sito scaricato nella PirateBox

Il sito riprodotto off-line con il metodo che abbiamo descritto si presta ad essere usato richiamandolo

con il comando **File -> Apri file…** del navigatore, non ad essere servito in un Web server, quale la PirateBox è. Se lo vuoi proporre attraverso la PirateBox, puoi fare quanto segue: comprimere in un file zip la cartella contenente il sito prodotta da WinHTTrack, caricare tale file zip fra i contenuti sulla penna USB della PirateBox, in modo che la gente possa fare il download del file zip e poi scompattarlo sul proprio device. Non è il massimo ma è quello che si può fare.

Ciò che ho scritto in questa sezione è ciò che ho ottenuto con le prove fatte fino ad ora, insoddisfacenti. Se scoprissi qualcosa di meglio o se qualcuno avesse qualcosa di meglio da suggerire aggiornerò.

# 3. Mi basta vedere la pagina web così com'è, senza pretendere di percorrere i link

## L'essenziale

Per fare questo basta utilizzare il comando **File -> Salva pagina con nome…** che si trova nel menu del navigatore. Il nome esatto del comando può variare un po' da un navigatore all'altro, ma è facilmente riconoscibile. Nella finestra che si apre, in basso a destra specificare l'opzione **Pagina web, completa**.

## L'approfondimento

Questa è la cosa più semplice da fare e probabilmente anche la più ragionevole. Deve essere chiaro però che così si scarica solo la pagina che si sta guardando. Quando si esegue questa operazione si creano due oggetti sul disco rigido: file HTML e una cartella. Per esempio se la eseguo su http:/iamarf.org, e chiedo di salvare con il nome **iamarf.html** [2] nella cartella **Scaricati**, in questa ci ritrovo un file di nome **iamarf.html** e una cartella di nome **iamarf_files**, che contiene tutti file grafici richiesti per la rappresentazione di quella pagina. L'occupazione totale, in questo esempio, è data da 161 KB del file HTML e 3.1 MB dei contenuti nella cartella associata. Una cosa del tutto ragionevole. Per vederla nel navigatore si carica il file **iamarf.html** con il comando **File -> Apri file…** La pagina appare correttamente, ma che succede se si cliccano i link? Dipende: se siamo online allora i link portano alle corrispondenti pagine presenti in Internet, perché questi rimangono inalterati nel processo di copia, mentre se siamo off-line allora non portano da nessuna parte, e si riceve un messaggio di pagina non trovata.

### Per offrire la pagina scaricata nella PirateBox

In questo caso è semplice, basta copiare i due oggetti, il file HTML e la cartella associata fra i contenuti nella penna USB della PirateBox e la pagina verrà servita regolarmente.

Se si prevede che possa interessante visualizzare anche alcune delle pagine che possono essere raggiunte a partire dai link, allora, in fase di preparazione dei materiali, si dovrà aver cura di scaricare separatamente anche queste.

## 4. Mi basta leggere i testi

Può succedere, che in certi casi una pagina contenga un testo di una certa lunghezza e che si sia interessati solo a questo. Ebbene, è qui che può tornare utile l'opzione **Pagina web, solo HTML…** del comando **File -> Salva pagina con nome…**. In questo modo ci ritroviamo solo un file in formato HTML, privato di tutta la grafica, che possiamo anche trasformare in qualche altro formato per leggerlo off-line in qualche altro modo.

[1] Questo risultato è stato ottenuto con una velocità di download di 53 Mbps, misurata con il servizio http://www.speedtest.net (http://www.speedtest.net).

[2] Salvando, il sistema propone di usare il titolo del blog come nome del file, ovvero **Laboratorio Online Permanente di Tecnologie Internet per la Scuola – #loptis.html** e per la cartella: **Laboratorio Online Permanente di Tecnologie Internet per la Scuola – #loptis_files**. Non è saggio usare nomi così lunghi e ingarbugliati. Specialmente gli spazi bianchi possono essere particolarmente malefici in alcune circostanze. Meglio optare per nomi più semplici.

[]