Générer une version statique d'un site Web

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 24 octobre 2008

http://www.bortzmeyer.org/generer-version-statique-site-web.html

---

On a souvent besoin de générer une version **statique** d'un site Web, c'est-à-dire de simples pages HTML, utilisables sans logiciel derrière, juste avec un serveur de fichiers, voire pas de serveur du tout, par exemple lorsque la version statique a été mise sur un CD-ROM. Mais comment faire avec les outils existants ?

Ces versions statiques sont très pratiques lorsqu'on veut pouvoir transformer un gros site Web nécessitant pour fonctionner un ensemble complexe de programmes en Java, en PHP ou autre, sans compter le SGBD qui l'accompagne. Cette transformation en site statique permet de consulter par la suite le site sur n'importe quelle plate-forme, même non connectée à l'Internet (ou bien mal connectée, un cas fréquent). `france.fr` aurait dû utiliser un tel mécanisme (puisque son contenu, quoique géré par des techniques dynamiques - Drupal - était entièrement statique), celà lui aurait épargné le ridicule.

Un autre cas qui m'a servi était celui où le moteur du site, un CMS, était bien vieux, plus maintenu, rempli de failles de sécurité et où personne n'avait le temps et l'envie de le mettre à jour. Le transformer en version statique a permis de continuer à le publier, même si on ne pouvait plus changer le contenu (<http://www.web1901.org/>). Cela avait l'avantage de supprimer beaucoup de risques de sécurité notamment les spam dans les commentaires.

Maintenant, comment faire pour produire cette version statique ? Si on a accès aux « cuisines », à la base de données où tout est stocké (ce qui est le cas des sites qu'on contrôle entièrement mais aussi de sites très ouverts comme Wikipédia <http://download.wikimedia.org/>), alors, il faut écrire un petit programme de conversion en HTML.

Si non, on peut toujours utiliser un client HTTP doté de la capacité de récupérer les pages, mais aussi de les modifier pour les adapter à cette nouvelle tâche. Il en existe deux en logiciel libre, wget et httrack.

Pour wget, l'utilisation de base est :

```
wget  --mirror --no-parent --convert-links http://www.LESITE.fr/
```

Le `--convert-links` est indispensable si certains liens sont absolus. Il faut alors les convertir en liens relatifs. Cette commande laisse un ensemble de fichiers HTML dans un répertoire nommé `www.LESITE.fr`, ensemble qu'on peut copier sur un CD-ROM ou une clé USB, archiver, etc.

Si le site est d'accès restreint, pas de problème :

```
wget --http-user MOI --http-passwd MONSECRET \
       --mirror --no-parent --convert-links  http://www.LESITE.fr/
```

wget a une limite que je trouve très gênante : si certains URL comportaient des ?, il laisse des fichiers avec un point d'interrogation dans le nom. Un navigateur comme Konqueror ne peut alors pas suivre les liens locaux (même sur Unix, il faut utiliser l'option `--restrict-file-names=windows` pour résoudre cela). Cela illustre l'importance de **tester** le résultat, surtout si on s'apprête à le graver sur un support stable.

Mais le vrai problème est que wget ne renomme pas les fichier en un nom finissant par `.html`. En local, le navigateur Web ne dispose pas de l'information donnée par le protocole HTTP et permettant de connaître le type du fichier récupéré. L'extension est donc indispensable (lynx avec `-force_html` ne résoud pas le problème car cette option n'agit que sur le **premier** fichier auquel on accède).

Le deuxième logiciel utilisable, et qui n'a pas ce défaut, est httrack. Si wget a beaucoup d'options (lisez son manuel en ligne), httrack en a une quantité astronomique. Mais on utilisation de base est aussi simple :

```
httrack http://www.LESITE.fr/
```

En outre, son affichage pendant l'exécution est bien plus agréable.

httrack réécrit les URL « dynamiques » en URL simples : `index.html?art=55` devient ainsi quelque chose comme `index6d76.html`. De tels fichiers sont bien plus facilement manipulables en local.

Les fichiers de httrack ne marchent toujours pas avec lynx (car les liens dans les fichiers ne sont pas toujours modifiés) mais c'est bon avec Konqueror qui gère intelligement les points d'interrogation. Ceci dit, l'option `-%q0` règle cela et j'utilise donc désormais httrack pour ces tâches.

Pour un autre article sur le même sujet, on peut consulter <`http://blog2doc.over-blog.com/article-1387761.html`>.