

Portabilité des données d'un site Web ?

Stéphane Bortzmeyer

<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 25 Octobre 2008

<http://www.bortzmeyer.org/portabilite-cms-blog.html>

Aujourd'hui, un grand nombre de sites Web sont techniquement gérés par un CMS ou un moteur de blog. Cela présente bien des avantages, notamment la séparation du **contenu** (placé dans la base de données) et de la **présentation**. Mais cela pose le problème de la **portabilité des données**. Si j'ai passé trois ans à remplir mon blog avec Wordpress, puis-je passer sur Dotclear ? Si j'ai tout mis dans SPIP, que se passe-t-il si je migre vers un site à base de Drupal ?

Avec les premiers sites Web, faits en HTML (à la main ou via un logiciel d'édition HTML comme nvu), c'était simple, HTML étant un format standard et ouvert (<http://www.bortzmeyer.org/formats-ouverts.html>), la portabilité des pages était automatiquement assuré. Mais HTML souffre d'autres inconvénients, notamment parce que la structure de la page (titre, menu, case pour le moteur de recherche, etc) et son contenu (spécifique à chaque page) sont mélangés. Si on veut mettre à gauche le menu qui était à droite, on n'a pas de solution simple, il faut reprendre toutes les pages. Donc, il est logique que des outils comme les CMS ou les moteurs de blog se soient imposés. Mais, à l'occasion, on a perdu la portabilité. Désormais, changer de logiciel devient difficile. Dans tous les cas, les gabarits sont perdus, parfois les informations sur les auteurs (qui peut écrire, qui peut modifier) mais peut-on au moins récupérer le contenu, ces articles qu'on a passé tant de temps à taper ? A priori, cela ne marchera pas : le schéma de la base de données est différent et le langage dans lequel sont décrites les données est également différent.

Cela a mené à la prise de conscience du problème de la **portabilité** des données, sous-ensemble du problème général de la préservation des ressources numériques. À l'heure actuelle, les utilisateurs sont, trop souvent, prisonniers d'un logiciel particulier (<http://www.mattcutts.com/blog/not-trapping-users-data-car>) en changer signifierait la perte d'un contenu significatif.

Commençons par les langages. La plupart des CMS ou moteurs de blog n'utilisent pas du tout HTML mais un langage de "*formatage*" formatage spécifique. Quels langages existent ? Textile, par exemple (Textpattern l'utilise), le langage de MediaWiki, rendu populaire par Wikipédia, des nouveaux formats normalisés comme Atom (RFC 4287¹), etc. Il n'y a rien qu'on puisse appeler « standard » dans

1. Pour voir le RFC de numéro NNN, <http://www.ietf.org/rfc/rfcNNN.txt>, par exemple <http://www.ietf.org/rfc/rfc4287.txt>

ce domaine. Rien que dans le monde du wiki, il existe de nombreux langages différents. Heureusement, ces langages sont en général plutôt simples, ce qui facilite l'écriture de programmes de conversion.

Ensuite, il y a la structure des données dans la base, différente d'un logiciel à l'autre, souvent mal documentée et toujours changeante.

Y a-t-il des programmes tout faits ? Oui, souvent, dès que le format qu'on veut importer est utilisé par un logiciel populaire. Par exemple, Wordpress dispose d'**importeurs** pour beaucoup de ses concurrents (http://codex.wordpress.org/Importing_Content). On peut les trouver sous l'onglet "Import" de l'interface d'administration. On peut lire de nombreux récits d'importation (<http://alexbrie.net/1504/how-to-import-textpattern-into-wordpress/>) de données... plus ou moins réussies. Wordpress lui-même dit prudemment que ces importeurs permettent de récupérer « la plupart » des informations associées aux articles de l'ancien blog.

En sens inverse, l'importance de Wordpress fait que beaucoup de moteurs de blog ont un moyen d'importer du Wordpress. Par exemple, pour Textpattern, on peut voir <http://forum.textpattern.com/viewtopic.php?id=23243> qui explique que "Click on administration->import, enter the database-information for the wordpress installation and click button. That will import the content.". De tels programmes d'importation "ad hoc" doivent être écrits pour chaque couple de CMS entre lesquels on veut échanger des données et il n'est donc pas étonnant qu'ils soient fragiles et ne respectent pas toujours toutes les données.

Et cela laisse ouvert le cas des logiciels « rares », pour lesquels de tels programmes ne sont pas disponibles.

Plusieurs personnes ont donc réfléchi à ce problème de **portabilité des données**. Ainsi, la création, en janvier 2008, du consortium DataPortability.org (<http://dataportability.org/>) par quelques acteurs de poids comme LinkedIn, Six Apart ou Flickr a suscité beaucoup d'intérêt (mais il semble qu'il n'y ai pas encore grand'chose de produit).

Quelques conseils pratiques, pour finir :

- Avant d'écrire des milliers de pages ou d'articles, demandez-vous comment vous les récupérerez en cas de changement de logiciel ?
- Si un programme de conversion semble disponible, testez-le d'abord, de préférence sur des données réelles. Beaucoup de ces programmes de conversion sont très sommaires et ne gèrent pas certains cas.

J'ai rassemblé quelques ressources Web sur ce thème en <http://del.icio.us/bortzmeyer/dataportability>.