Langages de balisage légers et logiciels de conversion de documents

Nicolas Poulain

28 mars 2012

Table des matières

1 Présentation

Pour saisir et mettre en forme des textes ou des documents textuels comportant des insertions d'images, de figures ou de tableaux, on utilise généralement un traitement de texte WYSIWYG [1], propriétaire comme Microsoft Word ou libre comme OpenOffice.

Les défauts majeurs de ces logiciels sont nombreux :

1. Le rédacteur d'un document se concentre presque autant autant sur le fond que sur la forme. Outre le temps passé, les conséquences sur le rendu sont nombreuses
   - Les mises en forme les plus hétéroclites sont autorisées au dépens de la lisibilité ;
   - Le résultat final est souvent discutable du point du vue de la typographie car les règles n'en sont pas respectées ni par l'utilisateur ni par le logiciel ;
   - L'utilisation des styles est souvent anarchique et les documents mal structurés, ce qui rend la production automatique de sommaire ou d'index impossible ;
   - L'insertion d'images ou de figures provoque des décalages mal maîtrisés.
2. En ce qui concerne les documents longs, l'inclusion de documents annexes au sein du document maître donne des résultats aléatoires ;
3. L'interopérabilité n'est pas assurée entre les logiciels, elle ne l'est pas même entre les différentes versions d'un même logiciel, ce qui nous amène au dernier point ;
4. La pérénnité des documents n'est pas certaine puisque la compatibilité ascendante ne fonctionne pas toujours et qu'un document écrit il y a quelques années risque d'être perdu, faute du logiciel capable de le lire.

À l'opposé de la composition dans un logiciel de traitement de texte, on peut écrire des documents dans des langages de balisage. Il en existe de nombreux : LaTeX, HTML, DocBook, etc. Les fichiers sont enregistrés au format texte brut et doivent être interprétés par un logiciel afin d'être consultés.

En ce qui concerne HTML et LaTeX, où pratiquement toutes les mises en formes sont possibles, le problème vient de la difficulté à écrire les balises [2]. Pour écrire un titre suivi d'une phrase contenant un mot en gras puis une liste non numérotée, on saisira respecivement :

– en LaTeX

```
\section{Le titre du paragraphe}

Voici un mot en \textbf{gras} puis une liste :
```

1. Un WYSIWYG pour *What you see is what you get* est une interface utilisateur qui permet de composer visuellement le résultat voulu. C'est une interface intuitive : l'utilisateur voit directement à l'écran à quoi ressemblera le résultat final.
2. Dans le cas du format DocBook, c'est même humainement presque impossible de l'écrire à la main tant l'enchevêtrement des balises est inextriquable. On le génère avec un logiciel WYSIWYG...

```
\begin{enumerate}
 \item  c'est simple ;
 \item  c'est efficace.
\end{enumerate}
```

– en HTML

```
<h1>Le titre du paragraphe</h1>

<p>Voici un mot en <strong>gras</strong> puis une liste :</p>

<ul>
 <li> c'est simple ;</li>
 <li> c'est efficace.</li>
</ul>
```

Comme on le voit, la syntaxe est accessible mais au goût de nombreux utilisateurs il y a trop de commandes de mise en forme qui nuisent à la lisibilité du texte lors de la saisie. C'est dommage car ces deux formats ouverts et universels ont chacun leur avantage :

– HTML peut être lu sur n'importe quelle plateforme ou terminal du monde entier car ses spécifications, gérées le W3C[3], sont respectées par les navigateurs web.
– le logiciel LaTeX produit des documents de qualité unanimement reconnue. Il prend en charge la mise en page, l'utilisateur n'ayant qu'à se concentrer sur le fond et sa structure.

Il existe une alternative qui est à la fois simple, interopérable et efficace : les langages de balisage légers.

Un langage de balisage léger est un langage utilisant une syntaxe simple, conçue pour qu'un fichier en ce langage soit aisé à saisir avec un éditeur de texte simple, et facile à lire dans sa forme non formatée.

Les wikis on grandement contribué à populariser ce type de langage. Le principe est de saisir des balises accessibles aux non inités, un moteur se chargeant de la conversion en HTML avant la publication.

```
Le titre du paragraphe
======================

Voici un mot en **gras** puis une liste :

* c'est simple ;
* c'est efficace.
```

Avantages :

– les balises sont visuelles et le texte reste lisible ;
– le nombre de balises et de règles à mémoriser est peu important ;
– les balises étant constituées de cractères non alphabétiques, on peut utiliser un correcteur d'othographe.

C'est en 1995 que l'on trouva la solution de ce problème, avec la création du premier langage Wiki, dont le but principal était de permettre l'édition facile de pages web par tout un chacun, et dont l'utilisateur actuel le plus célèbre est l'encyclopédie libre Wikipédia. S'il y a presque autant de syntaxes différentes que de logiciels Wiki, elles ont toutes la caractéristique d'utiliser des caractères textuels simples et intuitifs pour donner les indications de formatage du texte.

Toujours le même exemple, une nouvelle section en MediaWiki :

= Nouvelle section Wiki =

et une en Setext :

2 Nouvelle section Setext

Mais pourquoi limiter ces langages de balisage léger à la seule génération de HTML ? Pourquoi ne pas utiliser la même syntaxe pour différentes cibles (appelées backends, targets ou writers selon les logiciels), de manière

3. Un WYSIWYG pour *What you see is what you get* est une interface utilisateur qui permet de composer visuellement le résultat voulu. C'est une interface intuitive : l'utilisateur voit directement à l'écran à quoi ressemblera le résultat final.

à obtenir aussi bien une page web en HTML, qu'un document en LaTeX pour l'impression, ou qu'une page de man pour un logiciel ? Ce sont les logiciels qui poursuivent ce but qui m'intéressent, ils constituent pour moi l'avenir de la bureautique informatique, et j'ai été amené à les comparer pour en choisir un dans lequel m'investir comme développeur.

Pandoc est un logiciel de conversion de documents permettant, à partir d'un texte écrit dans un format très simple (Markdown), de faire une page HTML, un document PDF ou encore un texte au format MediaWiki par exemple.

L'auteur décrit son logiciel comme le couteau suisse de la création des documents. Ce logiciel permet de convertir des textes simples dans de nombreux formats, mais aussi de convertir différents formats entre eux, avec une qualité en constante amélioration. Plus besoin de logiciels lourds pour l'édition de vos documents : écrivez vos documents dans un simple éditeur de texte, et convertissez ensuite votre fichier dans le format de votre choix. Ce logiciel est idéal pour les documents structurés avec tableaux et images.