

Système automatisé de co-rédaction de livres

Yves Maniette

UNESP, Araraquara, São Paulo, Brésil

yves@iq.unesp.br

<http://sibee.iq.unesp.br/~yves>

Abstract

Producing a book that requires the collaboration of dozens or hundreds of people, as for a conference, is known to be a major challenge. Every organizer should be able to check every article, or every author be able to follow the relevant typographic rules or corresponding coding. On the other hand, the ubiquitous availability of ever-improving desktop publishing (DTP) programs has convinced everyone that their text is already “camera ready”, whereas most of the time, achieving good typography means that the typesetting has to be re-done, from scratch, thereby casting doubt on the validity of such DTP programs. In this paper we present a system that will perform the appropriate typesetting without any special typographic knowledge required of the writers. Thus, every author is able to use their everyday word-processing program to produce their text, because most of the detailed typographic work will be done silently, eventually resulting in a master file for the entire book, including any relevant indices.

Résumé

La rédaction d'un ouvrage produit en collaboration par plusieurs dizaines ou centaines de personnes, par exemple à l'occasion d'un congrès, n'est pas aisée. Elle exige des organisateurs les corrections requises dans chaque article, ou bien de chacun des rédacteurs le respect de règles précises de typographie ou de la codification correspondante. Par ailleurs, l'emploi généralisé de logiciels de mise en page chaque jour plus sophistiqués fait que nombreux sont ceux qui pensent produire un texte «bon à tirer», alors que bien souvent une typographie soignée exige que la mise en page soit entièrement refaite, ce qui met en cause le bien fondé de ces logiciels de mise en page. Nous présentons un système qui tient compte des nécessités d'une mise en page de qualité sans demander de grandes compétences aux rédacteurs, qui peuvent ainsi rédiger le texte de leur communication avec leur logiciel habituel. La plupart des traitements typographiques seront réalisés de manière transparente, jusqu'à produire le fichier maître du livre, y compris les annexes.

Introduction

De nombreuses sociétés savantes organisent leur congrès annuel par roulement : à tour de rôle, un groupe de membres se charge des multiples tâches administratives et techniques d'organisation du congrès. La plupart du temps, ce groupe a des connaissances et une disponibilité réduites. Une des tâches les plus difficiles à mettre en œuvre est justement la préparation du livre des actes. C'est sans doute la raison pour laquelle bien souvent les organisateurs de congrès optent pour la préparation d'un disque contenant les textes sous divers formats. Il n'est pas rare, toutefois, que lors de la première consultation du disque après la réunion on constate que certains fichiers ou illustrations ne s'ouvrent pas ou demeurent invisibles. Ce genre de disque contient aussi, bien souvent, un index automatisé qui permet par exemple de retrouver tous les textes contenant le mot «the» ou le caractère «;» ou «#» ! En revanche, il est parfois impossible d'y localiser un article avec plusieurs mots-clé, comme par exemple «cuivre», «protection» et «oxydation».

Pour de telles raisons, il est extrêmement important de pouvoir offrir à chaque participant un livre des actes du congrès dès le premier jour de la réunion. Idéalement, ce livre doit être typographiquement soigné et comprendre des index permettant de retrouver rapidement un texte quelconque ou l'adresse d'un correspondant.

Je me suis trouvé il y a un peu plus d'un an membre du groupe de préparation du Symposium brésilien d'électrochimie et électroanalytique (SIBEE), et bien que je ne sois ni chimiste ni Brésilien, je me suis spontanément proposé pour participer à la préparation du congrès et aider à l'élaboration des annales. C'est pourquoi je participe à la présente réunion, afin de faire part de mes réflexions.

Aspect logiciel et financier

Le budget initial du congrès étant modeste, il devait être possible d'en commencer l'organisation sans que cela n'entraîne de frais particuliers, ni en logiciels ni en matériel. Cette décision a aussi été motivée par la volonté

de pouvoir mettre à disposition d'autres groupes de recherche brésiliens un outil typographique et de gestion de documents qui soit librement utilisable. De plus, sachant que, hormis dans certaines grandes entreprises ou administrations, la plupart des ordinateurs fonctionnant au Brésil le font sans licence particulière, que les budgets alloués à l'achat de logiciels sont souvent dérisoires et que les logiciels libres y sont paradoxalement peu connus et inspirent la méfiance, il est apparu que l'emploi de logiciels libres s'imposait d'emblée.

Typographie et détails du livre

Le livre des actes devait contenir environ 300 documents de 2 à 3 pages contenant des équations mathématiques et chimiques, des références bibliographiques, des illustrations de tailles variables et parfois formées de plusieurs parties, ainsi que les index par auteurs, mots clés et titres. L'outil typographique devait permettre d'inclure aisément ces divers éléments, de classer les divers textes en temps réel jusqu'au dernier moment, quelques heures avant de transmettre le fichier final à l'imprimerie.

Les solutions existantes

Une recherche des produits existants a montré que n'existait guère que des logiciels fonctionnant sous licence d'exclusivité. De plus, la plupart du temps originaires des EUA, ceux-ci n'offraient que peu de libertés quant à la typographie ou le traitement des langues. Par exemple, les responsables de l'Initiative de Budapest pour le libre accès à la recherche ont publié un recensement des logiciels destinés à publier des journaux scientifiques¹. Toutefois, la plupart de ces programmes ne sont pas de code ouvert. Par ailleurs, un logiciel français employé dans de nombreux magazines en ligne, SPIP², a des atouts très intéressants mais n'offrait pas la composition des mathématiques ou de la chimie au moment où nous devons décider quel logiciel employer. SPIP a toutefois subi de notables améliorations très prometteuses.

En tout état de cause, j'ai créé un ensemble capable d'assurer au moins les fonctions suivantes :

- inscription des participants ;
- inscription de chaque article dans un chapitre déterminé ;
- réception des textes ;
- réception des illustrations ;
- traitement des illustrations en temps réel ;
- production de chaque article en temps réel ;
- aide à la typographie de la chimie ;
- production de transparents ou panneaux à partir du même document que l'article ;
- production du livre à tout moment.

1. Voir notamment <http://www.soros.org/openaccess/fr>
2. <http://www.spip.org>

Logiciels employés

Le serveur : logiciels employés en aval Le centre du système est un serveur programmé en script php qui fait appel à plusieurs logiciels, dont le principal est T_EX. Ghostscript est employé pour créer les fichiers au format pdf, et jpeg2ps³ pour transformer automatiquement en fichiers .ps les illustrations reçues au format jpeg.

Rédaction : logiciels employés en amont La grande majorité des participants au congrès, chimistes brésiliens, tout comme de nombreux scientifiques du monde entier, produisent leurs textes en employant un logiciel commercial fort connu dont le nom tient en un mot, et ne connaissent même pas d'autre logiciel. C'est dire que pour nombre d'entre eux, la question que pose Hans Hagen dans sa communication de ce congrès n'a même pas lieu d'être : au Brésil, de nombreuses sociétés savantes n'acceptent que les documents produits par «Word for Windows».

Recherche d'un compromis Il y a donc une apparente contradiction entre ma volonté d'employer des programmes qui soient en règle et accessibles à tous pour faire fonctionner le serveur, et la réalité des chercheurs attachés à un outil avec lequel ils produisent la majorité de leurs travaux mais dont les capacités typographiques ou de programmation sont limitées ; de plus, il est pratiquement inutilisable pour traiter de gros documents en toute sécurité.

Pour cette raison, le serveur devait pouvoir accepter des textes rédigés avec Word, mais pouvoir les faire traiter par T_EX. Pour ce faire, il a simplement été demandé à chaque participant de placer quelques codes simples dans le texte et de le sauvegarder au format de texte simple (.txt) avant de l'envoyer au serveur pour la compilation en direct.

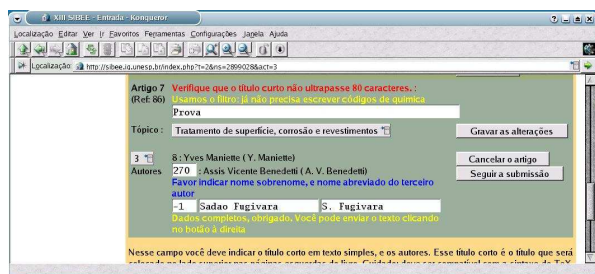
Commandes simplifiées Compte tenu de ce qui précède, des commandes extrêmement simples ont été créées, pour ne pas compliquer la tâche des rédacteurs. Elles comprenaient pour la plupart deux lettres majuscules et le symbole d'échappement. Par exemple, le titre de l'article est défini par :

```
\TR Titre du travail
```

suivi d'une ligne vierge. La macro utilise aussi d'autres variables créées automatiquement lors de l'enregistrement de l'article sur le serveur, qui définissent les noms des auteurs et de leurs institutions, ainsi que le titre court destiné à apparaître en haut des pages de gauche. Cette option est a priori dangereuse car les noms des auteurs ne sont pas enregistrés sur le même document que l'article. Toutefois, comme toutes les manipulations sont faites automatiquement, cette crainte n'est pas justifiée, et surtout,

3. Logiciel permettant d'enrober un fichier jpeg pour en faire un document PostScript, par Thomas Merz.

cette disposition permet d'éviter des erreurs sur les noms et d'aider à préparer l'index des auteurs. Présent ou non au congrès, chaque auteur doit s'inscrire sur le serveur, qui l'identifie de façon unique. On évite ainsi tout problème d'homonymie, et pour le responsable de l'article, il suffit d'indiquer les numéros d'inscription de chacun des co-auteurs pour qu'apparaissent automatiquement leurs noms et adresses sur le document. Voici par exemple la zone d'enregistrement d'un article :



Le nom du premier auteur ne peut pas être modifié, et les autres auteurs sont indiqués par leur numéro d'inscription, bien qu'il soit aussi possible de faire apparaître un auteur avec le numéro -1 si, par exemple, il ne juge pas utile d'apparaître dans l'index, et donc ne s'inscrit pas. En pareil cas l'auteur principal doit remplir les champs du nom et du nom abrégé.

Le résumé de l'article vient après le titre. Il est composé sur la largeur totale de la page, à la suite de quoi le système bascule automatiquement en mode de deux colonnes. Tout cela est réalisé par la commande

```
\SA Voici le résumé...
```

placée au début du texte et suivie d'une ligne vide. La macro copie aussi verbatim dans un fichier annexe le résumé («Sinopse do Artigo», en portugais). \TeX est alors prêt à recevoir le corps de l'article. De façon semblable :

```
\ABA Here is the Abstract...
```

placé n'importe où dans le texte et lui aussi suivi d'une ligne vierge copie le résumé en anglais dans un second fichier auxiliaire. Ce résumé sera composé sur une page unique placée à la fin de l'article, pour que le rédacteur puisse le relire. Les résumés en anglais devaient ensuite être exploités pour catalogage dans les *Chemical Abstracts*.

Une série de commandes comparables permet de créer des titres de paragraphe (\backslash TI), sous-paragraphe (\backslash TSI), listes numérotées ou alphabétiques, etc.

Un habitué de \LaTeX travaillant avec un éditeur adapté à ce logiciel pourra sourire en voyant de tels codes et pourra se demander pourquoi «réinventer la roue», car il lui est extrêmement facile d'écrire des commandes plus longues. En revanche, pour un usager de Word, de telles commandes pourtant simples sont déjà difficiles à lire, et l'option de correction automatique peut même perturber

le texte, car les noms de certaines macros ne signifient rien en portugais — ils sont «corrigés» automatiquement.

Langues La langue de base du congrès est le portugais mais souvent les références sont rédigées en anglais, ou en latin dans certains cas. Par ailleurs, certains participants venus d'Argentine ont rédigé leur communication en espagnol.

Le système a donc été compilé pour pouvoir traiter une dizaine de langues répondant aux cas de figures les plus probables, et une série de macros permet de basculer d'une langue à l'autre afin d'employer le dictionnaire de coupures adéquat. Enfin, certains textes composés automatiquement, comme par exemple les mots «figure» ou «tableau» ou «Références» sont définis en fonction de la langue de base de l'article, sans que l'auteur n'ait à s'en soucier.

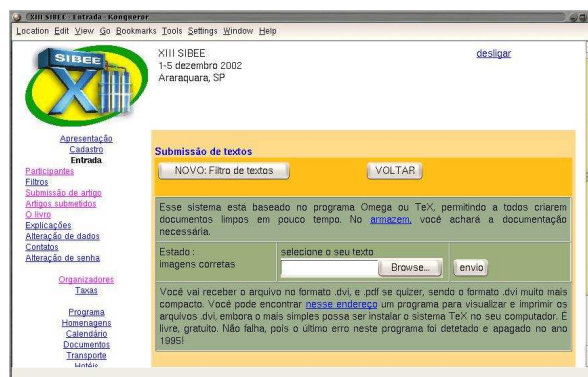
Illustrations Chaque illustration est indiquée par la commande suivante :

```
\FIG légende de la figure.[référence]
```

où la référence est un mot-clé employé par le système pour nommer le fichier de l'illustration correspondante, et qui permet aussi de définir la commande \backslash figure{référence} comme étant égale au numéro correspondant de figure.

Double compilation

Afin de pouvoir recevoir correctement les figures et de pouvoir y faire référence y compris avant leur apparition, chaque fichier est traité deux fois par le système, de manière transparente pour l'utilisateur. Avant la première passe, le visiteur doit simplement fournir au système le fichier du texte de son document :

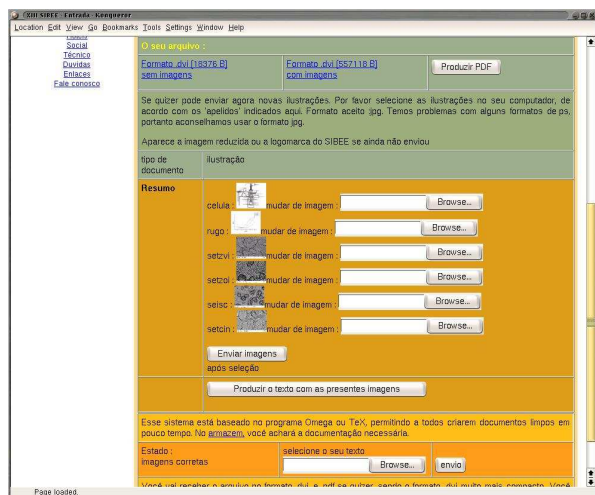


À ce moment, le script commande la compilation du fichier par \TeX , en utilisant une première série de commandes. Dans cette série, la plupart des commandes de mise en page sont réduites au strict minimum, mais les commandes de figures sont exploitées afin que \TeX prépare deux fichiers auxiliaires. Le premier contient les définitions lui permettant de connaître le numéro de chaque

figure, et le second est destiné à être de nouveau exploité par le script php. Ce second fichier comprend, pour chaque illustration, une ligne du type :
 $\$figr[1]="réf\grave{e}rence" ; ++\$totfigr ;$
 signifiant qu'à la variable $\$figr[1]$ on assigne la valeur "réf\grave{e}rence" et que l'on doit incrémenter le total du nombre de figures. Si les figures requises sont déjà présentes sur le serveur, la seconde compilation, menant à la production du fichier .dvi, est conduite automatiquement. En cas contraire le serveur présente une page telle que la suivante, montrant les réf\grave{e}rences des figures nécessaires, et un logotype «par défaut» quand elles sont absentes.



Chaque figure arrivant au serveur est transformée instantanément en fichier PostScript avec le programme jpeg2ps de Thomas Merz, et ce fichier PostScript est ensuite exploité immédiatement par Ghostscript pour former un second fichier jpg de basse résolution employé pour montrer l'onglet correspondant à l'auteur.



En tout état de cause, on peut à tout moment envoyer une nouvelle version du texte et de nouvelles ver-

sions des figures (au format jpeg). Les versions nouvelles du texte ou des figures viendront écraser les anciennes et les numéros de figures seront mis à jour instantanément. Aucune manipulation humaine n'est nécessaire et la présence de l'onglet sur le serveur prouve que le fichier pourra être compilé sans problème, au moins pour ce qui concerne les illustrations.

Chimie

Le paquet le mieux adapté qui permette d'écrire des symboles chimiques assez aisément est probablement ppchTeX, groupe de macros de ConTeXt qui est compatible avec T_EX et que l'on installe en chargeant les fichiers pictex.tex et m-ch-en.tex.

«Lavoisier»: le filtre Sachant que les textes sont envoyés au serveur sous la forme de texte selon le codage ISO 8859-1, on perd une quantité considérable d'information quant aux termes placés en italiques ou en exposants. Les commandes permettant de placer des italiques ont été indiquées aux auteurs et, dans un premier temps, une série de commandes correspondant à divers composés chimiques ou unités a été préparée. Cependant, les auteurs étant peu habitués aux textes à compiler, il est apparu qu'une solution plus simple et efficace était absolument nécessaire.

Un des aspects les plus ergonomiques du logiciel SPIP évoqué plus haut se trouve dans les «raccourcis SPIP» qui codifient simplement l'usage de l'italique, du gras ou la notation des liens hypertexte. Ces «raccourcis SPIP» sont destinés à être traités par un script php qui utilise des fonctions de remplacement automatique. Dans le présent système, un filtre semblable a été installé, qui permet de nombreuses transformations. Il doit encore voir de nombreuses améliorations mais la fonction a toutefois déjà reçu le nom pompeux de «Lavoisier».

CuSO₄ est ainsi transformé en $\backslash\{CuSO_4\}$, qui donne CuSO₄. Le code $\backslash q$ est un synonyme du code $\backslash chemical$ du paquet ConTeXt.

Un grand nombre de composés de chimie sont parfaitement descriptibles en format de texte ASCII car pour la plupart il existe une relation bijective entre le format de texte brut et le format composé selon les règles typographiques des chimistes.

En conséquence l'emploi d'un logiciel compliqué de mise en page est ainsi même remis en question puisqu'un tel filtre permet de s'affranchir des détails de typographie. En tout état de cause, il est toujours possible de bloquer le travail du filtre, par exemple en ajoutant des accolades que T_EX n'imprimera pas.

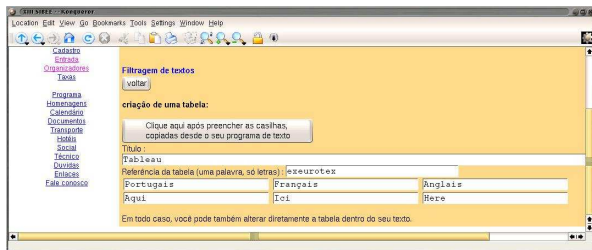
Le fichier de filtrage contient quelques centaines de lignes, et son exécution par le script php dure quelques dixièmes de seconde sur un texte de deux ou trois pages. Une seule passe permet de traiter tous les composés du

type $C_xH_yO_z$, par exemple, et quelques autres passes permettent aussi de traiter raisonnablement un grand nombre d'unités : $mV \cdot s^{-1}$ ou bien $mV \ s^{-1}$ sont alors transformés de suite en $mV \cdot s^{-1}$ qui donne donc $mV \cdot s^{-1}$ sur le papier. On notera au passage que ce dispositif permet d'assurer une bonne homogénéité des conventions typographiques, et qu'il peut être adapté, par exemple en forçant l'emploi du L majuscule pour le symbole du «litre» ou au contraire pour lui substituer le «dm³».

Le même travail est mené avec toutes les unités nécessaires, et peut être étendu à certains des codes typiques de \TeX que de nombreux auteurs oublient ou ne connaissent pas. Par exemple, le filtre ajoute automatiquement le caractère d'échappement avant %, \$, # ou &, mais sans en ajouter un second s'il y en a déjà un, et peut aussi éliminer des espaces parasites avant les virgules. Le traitement des espaces fines avant les symboles hauts de ponctuation sont en revanche traités directement par \TeX , car dans certains cas on veut pouvoir modifier le comportement du programme.

Programmé en script php, le filtre est modifiable à tout moment et peut aussi être mis à profit pour engendrer les codes html correspondants. Cela signifie que à partir d'un même document lisible par un être humain, nous pouvons créer un document \TeX ou un document destiné à être publié sur l'Internet.

Tableaux Une page spéciale de filtre a été préparée pour que chacun puisse rédiger sans trop de peine les codes permettant de créer un tableau. En premier lieu on indique le nombre de lignes et de colonnes. Alors le système présente un ensemble de cases qu'il suffit de remplir pour afficher ensuite le code complet correspondant au tableau. Ainsi la page suivante :



permet de préparer sur le champ les codes suivants, qui permettent de rendre le tableau :

```
\TV 3 colunas Tableau [exeurotex]
\PT | Portugais | Français | Anglais |
\NL | Portugais | Français | Anglais |
\NL | Aqui | Ici | Here | \FIMTAB
```

Mots-clés

Chaque article est susceptible de recevoir trois mots-clés, et dans l'index apparaîtront donc trois entrées par ar-

ticle, dans lesquelles les deux autres mots-clés seront indiqués entre parenthèse. On pourrait ainsi trouver rapidement dans le livre, parmi quinze articles traitant du carbone, ceux qui auraient rapport à sa protection contre l'oxydation. Cette façon de créer l'index des mots-clés peut paraître traditionnelle mais son simplicité apparente n'empêche qu'elle soit efficace et donne un document très compact.

Création du livre

À tout moment chaque participant au congrès peut afficher, en temps réel, les titres des articles en préparation, classés par chapitre. Les données sont stockées sur une base consultée à chaque demande du serveur, et cela ne surcharge pas excessivement la machine car les clients potentiels sont peu nombreux.

À tout moment également le responsable du site peut préparer un fichier du livre entier ou bien d'un ensemble d'articles, par exemple pour l'envoyer à l'un des membres du comité scientifique. Une fois de plus, c'est le script php qui prépare le fichier à faire compiler par \TeX . Les quelque 700 pages du livre des actes de ce congrès, contenant 265 travaux, sont compilées en moins d'une minute.

Améliorations nécessaires

Ce système a subi de nombreuses modifications au cours de son rodage, dont la plupart n'étaient pas imaginées au début de sa construction, car la différence énorme de vision entre un usager de programme de PAO grand public et celle d'un usager de \TeX est telle que la communication avec les participants au congrès fut parfois difficile.

Toutefois, chaque participant au congrès s'est accordé à reconnaître que le résultat obtenu est d'excellente qualité, et le livre des actes, finalement imprimé dans les 5 jours qui ont précédé le congrès, a fait l'unanimité des participants, et un système semblable sera probablement employé lors de la prochaine édition du SIBEE, en août 2004. Il reste néanmoins de nombreux détails à améliorer, tant du côté du serveur que du côté des programmes de rédaction.

En amont : rédaction Ma méconnaissance de certains outils du logiciel Word ou OpenOffice fait que je n'ai pas imaginé que l'on puisse créer certaines macros qui permettraient de sauvegarder le texte en un format qui soit acceptable par le système. Il existe, notamment disponible avec le logiciel Abiword, la possibilité de sauvegarder en format \LaTeX , mais ce format ne fait de reprendre trait pour trait la disposition du texte original tel qu'il apparaît lorsqu'il est mis en page par Abiword. Il faut au contraire extraire les données pertinentes, sans autre détail inutile.

En aval En aval, le serveur doit offrir des outils plus performants permettant de filtrer les textes de façon entièrement transparente ou de créer des tableaux complexes plus aisément. En particulier, une option mérite d'être sérieusement reconsidérée, qui consiste à employer \LaTeX au lieu de \TeX comme outil de mise en page, bien que la simplicité des commandes actuelles a rendu possible la création d'un ouvrage aussi volumineux, rédigé par une majorité de non spécialistes de \TeX . Le site du congrès est accessible à l'adresse <http://sibee.iq.unesp.br> où nous placerons une zone d'accès libre, et une fois qu'une version fonctionnant bien sera en service, elle sera mise à la disposition du public.

Remerciements

Erwin Bergamo a passé de nombreuses heures à corriger de nombreux textes ayant des problèmes de compilation. Durant trois mois il a fourni une aide précieuse, et pourtant il lui fallait simultanément terminer la rédaction de son mémoire de doctorat ; et son mémoire est un des premiers de notre institut à être mis en page avec \LaTeX .