Documentation et bibliothèques

De l'imprimé vers l'électronique : réflexions et solutions techniques pour une édition savante en transition

From Print to Electronic: Thoughts and Solutions for Scholarly Publishing in Transition

Del texto impreso al texto electrónico: reflexiones y soluciones técnicas para una edición académica en transición

Marie-Hélène Vézina and Martin Sévigny

Volume 45, Number 4, October–December 1999

Édition électronique

URI: https://id.erudit.org/iderudit/1032719ar
DOI: https://doi.org/10.7202/1032719ar

See table of contents

Publisher(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this article

Vézina, M.-H. & Sévigny, M. (1999). De l'imprimé vers l'électronique : réflexions et solutions techniques pour une édition savante en transition. *Documentation et bibliothèques*, *45*(4), 161–172. https://doi.org/10.7202/1032719ar

Article abstract

Publishers and librarians, the intermediaries in the document chain, are increasingly involved in the electronic dissemination of information projects. At the same time, they continue to meet the reader's need for printed material. The different approaches used to insure this co-habitation can be summarised as follows: (1) produce the printed copy then generate electronic versions, (2) produce both formats at the same time, and (3) produce both formats using a single source document containing all the semantic information required for both operations. This article discusses the main advantages of the third option and describes an application developed in pilot project of scholarly journals at the Presses de l'Université de Montréal. The authors discuss the choice of dissemination formats and storage. The SGML format was chosen because of its ability to integrate, its durability, and the semantic richness it expresses. The pilot project (Érudit) described in this article consists in developing a highly automated chain of operations using SGML, a format that will soon be replaced by the XML format. The SGML format will automatically generate the HTML versions, a format widely used on the Web, as well as the print formats, such as PostScript and PDF, used respectively for printing on paper and printing from a distance.

De l'imprimé vers l'électronique : réflexions et solutions techniques pour une édition savante en transition

Marie-Hélène Vézina
Chargée de projet - Édition électronique
Presses de l'Université de Montréal
marie-helene.vezina@umontreal.ca
Martin Sévigny
Chargé de projet — Édition électronique
Presses de l'Université de Montréal
sevigny@ajlsm.com

Les acteurs intermédiaires de la chaîne documentaire, soit les éditeurs et bibliothécaires, sont de plus en plus impliqués dans des projets de diffusion électronique d'information. En même temps, ils doivent continuer à répondre à la demande de supports imprimés. Les différentes approches pour assurer cette coexistence se résument aux trois scénarios suivants : produire la forme imprimée pour en dériver par la suite des versions électroniques ; produire parallèlement les deux formes ; ou enfin, produire les deux formes à partir d'un document source unique contenant toute information sémantique requise pour ces deux opérations. Cet article étudie les principaux avantages de cette troisième approche et en présente une application dans le cadre d'un projet pilote de revues savantes aux Presses de l'Université de Montréal. Les auteurs abordent les questions entourant le choix des formats de diffusion et d'archivage. Le format SGML a été retenu pour ses qualités d'intégration, sa pérennité et la richesse sémantique qu'il peut exprimer. Érudit, le projet pilote présenté en détail ici, a consisté à développer une chaîne de traitement fortement automatisée basée sur le SGML, lequel format sera bientôt remplacé par le format XML. Du format SGML sont produits automatiquement des versions HTML, format standard de diffusion sur le Web, ainsi que des formats d'impression, soit PostScript et PDF, respectivement destinés à l'impression de la publication papier et l'impression à distance.

From Print to Electronic : Thoughts and Solutions for Scholarly Publishing in Transition

Publishers and librarians, the intermediaries in the document chain, are increasingly involved in the electronic dissemination of information projects. At the same time, they continue to meet the reader's need for printed material. The different approaches used to insure this co-habitation can be summarised as follows : (1) produce the printed copy then generate electronic versions, (2) produce both formats at the same time, and (3) produce both formats using a single source document containing all the semantic information required for both operations. This article discusses the main advantages of the third option and describes an application developed in pilot project of scholarly journals at the Presses de l'Université de Montréal. The authors discuss the choice of dissemination formats and storage. The SGML format was chosen because of its ability to integrate, its durability, and the semantic richness it expresses. The pilot project (Érudit) described in this article consists in developing a highly automated chain of operations using SGML, a format that will soon be replaced by the XML format. The SGML format will automatically generate the HTML versions, a format widely used on the Web, as well as the print formats, such as PostScript and PDF, used respectively for printing on paper and printing from a distance.

Del texto impreso al texto electrónico : reflexiones y soluciones técnicas para una edición académica en transición

Los intermediarios de la cadena documental, es decir, los editores y los bibliotecarios, se ven cada vez más envueltos en proyectos de difusión electrónica de información. Al mismo tiempo deben seguir respondiendo a la demanda de lectores de materiales impresos. Los diferentes enfoques para asegurar esta coexistencia se resumen en las tres situaciones siguientes : producir material impreso para derivar del mismo versiones electrónicas ; producir paralelamente los dos formatos ; o, finalmente, producir los dos formatos a partir de un documento original único con toda la información semántica necesaria para estas dos operaciones. Este artículo trata las principales ventajas relacionadas con el tercer enfoque y presenta una aplicación del mismo dentro del marco de un proyecto piloto de revistas académicas en la Editorial de la Universidad de Montreal. Los autores abordan asuntos relacionados con la elección de formatos de difusión y archivado. Se eligió el formato SGML por sus cualidades de integración, perennidad y riqueza semántica que puede expresar. El proyecto piloto (Érudit), presentado en detalle, consistió en elaborar una cadena de tratamiento altamente automatizada y basada en SGML, que se reemplazará muy pronto por el formato XML. A partir del formato SGML se producen automáticamente versiones en HTML, formato corriente de difusión en la Web, así como los formatos de impresión, como PostScript y PDF, destinados, respectivamente, a la impresión convencional y teleimpresión.

De nombreux « systèmes d'information » sont mis en place dans le but d'atteindre un objectif en apparence assez simple : favoriser la circulation de l'information depuis les créateurs de celle-ci (les auteurs), jusqu'aux consommateurs (les lecteurs), afin que ces derniers puissent en profiter et satisfaire leurs besoins en information. Ces systèmes, que l'on a cherché depuis le début à rendre de plus en plus efficaces, sont à la fois basés sur des interventions humaines et des technologies. Aujourd'hui, ce sont les réseaux informatiques, pour ne pas dire Internet, qui semblent dicter la voie à suivre en matière de systèmes d'information.

Bien entendu, tous les acteurs doivent s'adapter à ces nouvelles pratiques et ces nouveaux outils. Parmi ces acteurs, on retrouve les bibliothèques et les éditeurs (Odlyzko 1999), intervenants qui se situent habituellement au centre du système d'information, quelque part entre les producteurs et les consommateurs. Ils doivent entre autres choisir, assembler, répertorier, diffuser, conserver l'information et, pour ce faire, ils doivent utiliser les systèmes – toujours technologiques et humains – les plus efficaces possibles.

Dans cet article, nous présentons comment un éditeur, en particulier un éditeur de revues savantes, peut s'y prendre pour exploiter les nouvelles technologies et ainsi rendre de meilleurs services aux bibliothèques et aux lecteurs, tout en conservant des coûts de production suffisamment bas pour demeurer compétitif. Dans une première partie, nous présenterons brièvement différents modèles d'édition électronique et les éléments à considérer pour les évaluer. Dans la seconde partie, nous décrirons une application réelle de l'un de ces modèles, issue des travaux en édition électronique réalisés dans le cadre du projet Érudit[1] aux Presses de l'Université de Montréal (Boismenu *et al.* 1999).

Cette application, et par le fait même cet article, ne prétendent pas résoudre tous les problèmes ni explorer tous les enjeux reliés à l'édition électronique. En fait, le projet fut réalisé dans un contexte particulier qu'il est nécessaire de rappeler ici. Le but du projet était de mettre en place un centre de service pour l'édition électronique de revues savantes québécoises. Les services sont offerts à tous les éditeurs intéressés de publier leurs revues *à la fois en format électronique et en format imprimé.*

Ce centre de service joue donc un rôle très important dans le contexte actuel de l'édition savante. En effet, la plupart des éditeurs ou des comités de rédaction de revues savantes expriment le souhait qu'elles soient diffusées sur Internet, tout en conservant des versions imprimées. Bien que l'objectif ici ne soit pas de présenter les avantages et inconvénients des formats électronique et imprimé, rappelons que la plupart des intervenants de la chaîne documentaire (auteurs, éditeurs, bibliothécaires et lecteurs) s'entendent sur les points suivants : 1) le format imprimé est encore utile ; 2) le format électronique est maintenant essentiel ; 3) de nouveaux « services aux utilisateurs » doivent être proposés à partir du format électronique ; 4) offrir une version imprimée sera important pour une période inconnue de façon générale, mais probablement encore assez longue, sinon éternelle.

Un éditeur désirant tenir compte de ces conclusions fait donc face à un double défi : offrir aux lecteurs des produits électroniques novateurs et utiles, tout en maintenant une version imprimée de bonne qualité. Inutile de préciser que les coûts de production doivent être maintenus à des niveaux très bas, surtout dans le domaine de l'édition savante en sciences humaines et sociales.

Ce contexte est situé au cœur des réflexions, applications et conclusions contenues dans cet article. En effet, nous cherchons d'abord et avant tout à proposer des solutions pour une période de transition où les versions électroniques et imprimées de revues savantes se côtoient et se complètent. Bien sûr, plusieurs éléments discutés s'appliquent à d'autres contextes, par exemple pour d'autres types de document ou encore pour des publications purement électroniques. Toutefois, nous n'aborderons pas ici ces derniers cas de figure.

Modèles techniques de publication

Au cours des dernières années, les réseaux, et en particulier Internet, nous ont permis de réaliser de nombreux projets de « diffusion électronique d'information », c'est-à-dire d'utilisation du support électronique pour la production et surtout la diffusion d'information. Plusieurs techniques nous permettent d'y arriver, et presque toutes ont comme point commun l'utilisation des formats de documents popularisés par Internet, soit HTML[2] et PDF[3]. Ces initiatives ont eu pour effet de constituer un immense réservoir d'information en format électronique, soit le World Wide Web, réservoir intéressant mais difficile à exploiter. Les éditeurs, pour qui la diffusion d'information n'a plus de secrets depuis fort longtemps, se sont bien entendu lancés dans cette grande aventure, la plupart embrassant les techniques habituelles : production et diffusion de HTML sur le Web ou encore production de documents PDF à partir de leur chaîne de traitement traditionnelle et diffusion de ces documents sur le Web.

Mais ces techniques sont-elles suffisantes pour assurer une édition de qualité, en particulier dans le monde de l'édition savante ? Les prochaines sections sont consacrées à cette question et passent en revue certains critères importants, en plus d'expliquer les différentes techniques utilisées pour chaque aspect de l'édition scientifique.

Moyens de production

Au cours des 20 ou 30 dernières années, les éditeurs ont su profiter des développements technologiques dans le domaine de l'informatique. Ils ont utilisé l'informatique dans la chaîne de production, à l'aide des techniques de publication assistée par ordinateur (PAO), que ce soit au moyen de logiciels de traitement de texte, de graphisme ou de mise en page. D'ailleurs, la plupart des éditeurs fonctionnent toujours avec ces moyens de production. Cette intégration des technologies n'est toutefois pas complète ni idéale. Ainsi, les derniers détails de l'impression sont très souvent ajustés de manière non informatique, par exemple le traitement de certaines images en procédé photo, le montage de différents fichiers en un document

1. Pour en savoir plus sur le projet Érudit et pour consulter le rapport complet sur le projet, voir <URL :http ://www.erudit.org/>
2. *Hypertext Markup Language*, norme du World Wide Web Consortium. Voir <URL : http ://www.w3.org/MarkUp/>.
3. Le format PDF (*Portable Document Format*) est un format de document électronique développé par la compagnie Adobe. Un document PDF conserve l'allure originale de la forme imprimée du document (textes, graphiques, couleurs) peu importe la plateforme utilisée.

continu ou l'ajout de pages disparates (encarts, annexes). De plus, même si ces chaînes de traitement utilisent massivement l'informatique, leur finalité est de produire des documents imprimés, et non des documents électroniques.

Nous pouvons tirer de ces méthodes deux conclusions en apparence contradictoires : presque tous les documents imprimés, même ceux produits de façon traditionnelle, existent sous une forme électronique quelconque, mais pour plusieurs documents imprimés, nous n'avons pas de version *finale et définitive* en format électronique. Par conséquent, même si en apparence les outils de PAO nous permettent de faire un pas vers de la véritable édition électronique, des ajustements doivent être faits afin d'obtenir un document électronique qui soit fidèle à la version imprimée quant au contenu. Cette utilisation de la PAO n'est donc pas suffisante pour obtenir un document électronique de qualité et pour assurer la diffusion ainsi que la conservation sur support électronique. Cette approche est donc nettement insuffisante, ce qui demande aux éditeurs de remettre en question non seulement leurs façons de diffuser et de conserver leur information, mais également de la produire.

Cette courte réflexion sur les méthodes actuelles nous amène à identifier trois grands scénarios pour la production de documents électroniques dans le monde de l'édition :

- *La forme électronique en aval de la forme imprimée.* Conserver les techniques de PAO traditionnelles et, à partir des résultats de celles-ci, créer des versions électroniques des documents ;
- *Les chaînes parallèles.* Continuer à travailler avec les outils de PAO pour les versions imprimées, mais « remonter » à la source (par exemple, des documents de traitement de texte produits par les auteurs) pour créer des versions électroniques ;
- *Un seul document source et des produits dérivés.* Produire d'abord un document riche et, par la suite, dériver des produits d'information, y compris l'utilisation de la PAO pour des versions imprimées.

La première approche souffre de deux lacunes importantes. D'abord, elle ajoute des étapes à la chaîne de traitement, ce qui la rend nécessairement plus coûteuse que le modèle traditionnel où l'imprimé constituait l'unique produit. De plus, puisque la chaîne de traitement est d'abord et avant tout orientée vers l'imprimé, il sera difficile d'exploiter les possibilités des documents électroniques.

La deuxième approche permet, peut-être, d'exploiter les possibilités des documents électroniques, mais encore une fois elle vient ajouter des étapes à la chaîne de traitement, ce qui la rend nécessairement plus coûteuse. De plus, une autre difficulté s'ajoute, car le fait de mener deux chaînes de traitement en parallèle rend les étapes de correction plus difficiles et les possibilités d'erreurs plus probables.

La troisième approche peut s'avérer intéressante et surtout efficace, puisqu'une seule chaîne de traitement est utilisée et qu'il n'y a donc pas de répétition de l'information. De plus, elle peut utiliser toute la puissance des outils de PAO pour produire des versions imprimées des documents, si nécessaire. Mais cette approche est possible seulement si on arrive à produire ce document riche à partir duquel les autres formats seront dérivés. Ces produits dérivés pourraient être, par exemple, une version imprimée et reliée, un document HTML sur le Web, une version sommaire de l'article (titre, auteurs, résumé) envoyée par courrier électronique à une liste de diffusion, etc.

Pour arriver à implanter un tel modèle de traitement et de production, il est nécessaire d'utiliser une technologie qui permet de créer des documents suffisamment riches pour représenter toutes les informations nécessaires aux traitements à effectuer, immédiatement ou dans les années à venir. Heureusement, une telle technologie existe, et il s'agit de la norme SGML (*Standard Generalized Markup Language*, ISO 8879), et de sa cousine XML (*Extensible Markup Language,* recommandation du World Wide Web Consortium[4]). Ces normes permettent de créer des documents structurés, c'est-à-dire des documents (nécessairement électroniques) qui contiennent de l'information à propos de leur contenu et de leur structure, plutôt que des informations de formatage en fonction d'un contexte particulier. Par exemple, pour un article scientifique, un document structuré contiendra de l'information sur la signification de ses différentes parties, par exemple un titre, un chapitre, un auteur, une référence bibliographique, etc.

À partir de ce document structuré, il est possible de dériver différents produits, car ceux-ci contiennent moins d'information que le document structuré, ou encore ils contiennent des renseignements plus détaillées mais faciles à déduire. Par exemple, il est facile de convertir une information telle que « ceci est un titre de section » en une série d'instructions de formatage telles que « mettre en caractères gras, 12 points, police Arial, espace de 12 points avant le paragraphe ». Il faut noter, et c'est important, que l'inverse n'est pas vrai : les instructions de formatage peuvent difficilement être converties en informations sur la structure, sauf si elles sont très exclusives et surtout très cohérentes à l'intérieur du document. Il se peut également que certains composants n'affichent aucun attribut de formatage particulier mais qu'il existe des besoins où cette information doive être distinguée, par exemple le nom de l'organisme d'appartenance d'un auteur dans le cas où on désirerait recycler cette information pour l'inclure dans un carnet d'adresses ou une liste d'étiquettes postales.

Ce dernier modèle de production, centré sur l'exploitation du document structuré, est fondamentalement différent du modèle de la PAO traditionnelle en ce qu'il nous permet de considérer les différents supports ou formats de diffusion, ainsi que les différents formats de conservation, comme étant des produits dérivés à partir d'une même source. Cette dérivation est en général assez aisée, et surtout elle peut être automatisée.

Dans le cadre du projet Érudit, nous avons implanté une chaîne de traitement basée sur une telle approche dont la méthodologie sera décrite en détails dans la deuxième partie de cet article.

Formats de diffusion et de conservation

Tout document électronique sera représenté par un fichier informatique, mais le contenu exact de ce fichier sera déterminé par le format utilisé pour représenter l'information. Très souvent, les formats de documents sont associés à l'application qui produit le document, par exemple les

4. Voir <http://www.w3.org/TR/1998/REC-xml-19980210>

formats *Excel* ou *WordPerfect* (pour une discussion exhaustive sur la question des formats de documents électroniques, voir Marcoux 1994).

La question des formats est probablement la plus importante dans un contexte d'édition savante électronique. Elle a des impacts majeurs sur la production, la diffusion ainsi que la conservation des documents.

Conservation

Puisque les documents électroniques doivent être stockés dans un format donné, il est nécessaire de s'interroger sur les critères à utiliser dans le choix d'un format de document électronique, en ayant pour objectif la conservation à long terme (voir une discussion intéressante à ce sujet dans Bullock 1999).

D'abord, il faut que le format soit capable de représenter correctement l'information contenue dans le document. Par exemple, s'il y a du texte et des images, il est nécessaire d'utiliser un format qui permette d'intégrer à la fois des informations textuelles et graphiques. De nos jours, ce n'est plus un véritable problème, car la plupart des formats de document permettent d'intégrer différents types d'information et sont à proprement parler des formats de documents « multimédias ».

Nous voulons également un format qui puisse être « lu » par une application, et ce, aussi longtemps que nous le souhaitons. C'est à ce stade-ci que les difficultés se présentent habituellement ; les conversions d'un format à l'autre (par exemple *PageMaker* à *Word*) ou encore d'une version à l'autre d'un même format (par exemple *WordPerfect 5.0* à *WordPerfect 8*) ne sont pas une solution à ce problème puisque plus souvent qu'autrement, il y a des pertes d'information, des changements dans la présentation et d'autres manifestations indésirables.

Heureusement, il existe une façon intéressante et éprouvée de contourner ce problème. Il s'agit d'utiliser à la fois un format de représentation de l'information très simple et très universel, et d'utiliser une technique qui rende ces documents « lisibles par l'humain ». Un document structuré représenté à l'aide de XML est un bel exemple d'un tel document.

En effet, on peut représenter un document XML à l'aide de caractères faisant partie du jeu de caractères ASCII. Concrètement, le fichier produit sera un pur fichier ASCII[5], soit le type de fichiers le plus universel que l'on trouve dans le monde informatique. Il y a de fortes chances qu'il existera encore des plates-formes informatiques et des applications qui nous permettront de « voir » un fichier ASCII, et ce, pour encore une très longue période. De plus, un document structuré utilisant XML contient de l'information du genre « <titre>Introduction</titre> ». Même sans application particulière, même sans connaissances informatiques ou de XML, il est assez facile de s'imaginer que le mot « Introduction » constitue ici un titre, et non le nom d'un auteur.

Diffusion

Les formats de diffusion de documents électroniques sont multiples et variés. Toutefois, il y a présentement une assez forte convergence vers deux formats associés au Web : HTML et PDF (voir à ce sujet Lieb 1999). Le format PDF est particulièrement bien adapté pour la représentation exacte de documents imprimés dans un format facilement diffusible sur le Web, car le logiciel pour consulter les documents PDF est gratuit, disponible en plusieurs langues et bien connu des utilisateurs. De plus, il est facile de produire des documents PDF à partir de n'importe quelle application informatique. Il est donc tout à fait naturel d'utiliser ce format à des fins d'impression à distance et sur demande, ce qui répond à de nombreux besoins pour l'édition savante, qui a souvent de faibles tirages ou encore parce que les usagers n'ont besoin que d'une partie des documents (un article plutôt qu'un numéro d'une revue, par exemple).

L'autre format de convergence est évidemment le format HTML, né avec le Web et popularisé avec l'évolution de ce réseau. Presque toutes les applications documentaires peuvent maintenant lire ou produire des documents HTML. Lorsqu'on associe HTML avec le langage JavaScript et les feuilles de style CSS[6] et qu'on obtient ainsi du Dynamic HTML (DHTML), il est possible de créer de véritables interfaces de consultation et non de simples documents électroniques (Dugand-Saenz et Verdret 1998). HTML est donc un excellent format de diffusion, mais malheureusement trop pauvre pour la gestion ou la production de documents. De plus, pour produire un document HTML de qualité, c'est-à-dire une interface de qualité, on doit travailler à partir d'une source d'information très riche, sinon le travail devra être fait à la main et sera fastidieux.

Ces deux remarques nous amènent à conclure que le format HTML s'avère un format de diffusion à privilégier, en autant que l'on utilise un autre format de gestion et que l'on puisse produire facilement des documents HTML de qualité. Ajoutons que les formats de PAO ne remplissent ni l'une ni l'autre de ces conditions, mais que les normes SGML et XML, elles, satisfont ces besoins.

Format de production ou de gestion

Au moment de la production de l'information, nous devons travailler avec un format qui nous permette d'atteindre tous les objectifs fixés dans les sections précédentes, et ce de façon efficace. En résumé, nous recherchons un format d'encodage de l'information qui nous permette de répondre à nos besoins, soit :

- manipuler aisément les documents pour effectuer toutes les activités de production (gestion, *workflow*, diffusion, etc.) ;
- permettre l'exploitation de toutes les possibilités qu'offrent les documents électroniques (multimédia, hypertexte, génération dynamique de contenu, recherche plein texte, données complémentaires/supplémentaires, etc.) ;
- produire des documents électroniques dans d'autres formats (par exemple HTML), et ce, en exploitant toutes les possibilités de ces formats ;
- permettre la diffusion sur différents supports (cédérom, DVD, réseaux, etc), y compris le support imprimé à l'aide d'applications de PAO ;
- conserver à long terme et dans des conditions optimales l'information et sa structure afin d'en assurer la pérennité.

Les documents structurés constituent la meilleure façon de répondre efficacement à l'ensemble de ces critères. À l'opposé, les formats associés à la PAO n'offrent pas la même polyvalence ni la même puissance, car ils contiennent de l'information en fonction d'un seul et unique support.

5. Techniquement, les documents XML sont stockés en Unicode, mais on peut les réduire à du simple ASCII sans perte d'information.
6. Cascading style sheet, voir <http://www.w3.org/Style/css/>

Exploitation du format électronique

Passer de l'imprimé à l'électronique constitue un changement qui va bien au-delà du mode de diffusion d'une revue savante. Les formats électroniques permettent en effet de représenter plusieurs types d'information que l'on ne peut retrouver dans un document imprimé. Une revue en transition vers l'électronique intégrera peu à peu ces types d'information et son éditeur devra mettre en place les outils nécessaires pour y arriver.

L'**information statique** est la plus évidente, mais aussi la seule qui peut être véritablement représentée sur une feuille de papier. Il s'agit de textes ou d'images qui, une fois « imprimés » ou « stockés » dans le document, ne changeront pas. La plupart des documents existants ne contiennent que de l'information statique, car ils ont été produits d'abord et avant tout pour un support qui ne permet que ce genre d'information, soit l'imprimé.

L'**information dynamique** est celle qui « bouge », qui s'anime. Ces animations ne sont pas contrôlées par les utilisateurs (ou si peu), mais plutôt par les producteurs de l'information. Le meilleur exemple est la vidéo ou les images en mouvement. La plupart du temps, l'interaction de l'utilisateur se limite à des fonctions telles que « marche avant » ou « arrière », « pause », « arrêt », etc. Les séquences sonores font également partie de ces informations dynamiques.

L'**information interactive** est celle qui peut prendre différentes formes ou valeurs en fonction du désir de l'utilisateur. Elle se distingue de l'information dynamique par l'importance qu'elle accorde au contrôle par l'utilisateur. Par exemple, la simulation d'une molécule en trois dimensions, avec la possibilité pour l'utilisateur de manipuler la molécule dans tous les sens pour la voir sous tous ses aspects, constitue de l'information fortement interactive. Un autre exemple consiste en la publication d'un algorithme auquel l'utilisateur peut fournir des valeurs de départ et vérifier les résultats, et ce, de façon instantanée ou presque. Cela peut aller d'une simple calculatrice d'intérêts composés à la simulation de la puissance d'un moteur.

L'**information active** permet aux utilisateurs d'agir sur le contenu du document ou encore sur l'environnement de consultation. Les liens hypertextuels font partie de cette catégorie, de même que les formulaires interactifs. Par exemple, un sondage publié dans un article scientifique pourrait être mis à jour dynamiquement par des lecteurs qui pourraient faire connaître leur opinion à partir du document.

Un des problèmes majeurs pour les éditeurs et, par le fait même, pour les utilisateurs est l'absence ou la surabondance de normes pour certains types d'information. Dans le cas du texte et des images simples, la situation est assez facile à maîtriser de par l'omniprésence du format HTML et de ses formats d'image associés, GIF et JPEG. Mais au-delà de ces quelques formats, la situation devient plus difficile, car le support n'est habituellement pas inclus dans les navigateurs communs, et il faut donc inciter les utilisateurs à installer des modules externes ou des applications supplémentaires afin de pouvoir consulter certaines parties de documents. Un utilisateur sera enclin à installer un tel module s'il en a besoin au moins occasionnellement ou si l'information manquante est très importante pour lui. Sinon, il va aller voir ailleurs ou il s'en passera.

Le défi technologique est double : trouver des formats adéquats pour chaque type d'information susceptible de se présenter et trouver un format de base qui puisse lier tous ces types d'information et qui serve de « ciment » aux différentes parties du document électronique.

Dans le cas du format de base, un modèle de traitement centré sur XML peut s'avérer suffisant. En effet, XML permet d'intégrer des parties de documents en différents formats. Ainsi, assembler un document ayant des composantes textuelles, iconographiques, vidéo, sonores et des algorithmes n'est pas un réel problème. De plus, si on utilise HTML comme principal format de diffusion pour la consultation électronique de l'information, nous avons là également un format qui peut assembler des documents très complexes comprenant des parties très différentes et stockés dans des formats variés. Bref, XML et HTML sont tous deux des formats « hypermédias » et ils constituent des solutions intéressantes pour la gestion et la diffusion de tels types de documents.

L'autre partie du défi est plus problématique : quel(s) format(s) utiliser pour les différents types d'information ? À ce sujet, un grand effort de normalisation reste à faire avant de s'assurer que les navigateurs habituels puissent présenter tous les types d'information sur toutes les plates-formes. Toutefois, soulignons certains aspects encourageants, tels que la mise en place d'une norme de l'industrie pour les images vectorielles (SVG[7]) et l'utilisation croissante du langage de programmation Java pour les applications dynamiques (ce qui pourrait être utile pour les algorithmes et les simulations).

Aujourd'hui, un éditeur qui prend au sérieux l'édition électronique et l'exploitation optimale des possibilités qu'offrent les documents électroniques devrait entreprendre ces différentes démarches :

- *sensibiliser, instruire les auteurs* potentiels aux possibilités des documents électroniques. Sans matière première, il est inutile de mettre en place des systèmes sophistiqués ;
- *sensibiliser les utilisateurs* aux possibilités des documents électroniques. Sans demande, l'offre ne sera pas nécessaire ;
- *identifier les formats de diffusion* les plus adéquats. Il y a deux questions fondamentales à se poser. Est-ce que le format choisi permettra de représenter adéquatement l'information à diffuser ? Est-ce que les utilisateurs possèdent les équipements et logiciels nécessaires pour pouvoir consulter des documents utilisant ce format ?
- établir des *protocoles pour l'échange de tels documents*. Les auteurs et les éditeurs doivent être en mesure de se transmettre efficacement ce genre d'information ;
- *mettre en place une infrastructure de gestion* pour ces types de documents. L'éditeur doit être en mesure de manipuler et gérer ces parties de documents et, bien souvent, les formats de diffusion et/ou d'échange ne sont pas les meilleurs pour y arriver, surtout si l'on considère la nécessité de conservation à long terme, par exemple ;
- *participer aux efforts de normalisation* des formats et des applications. Les éditeurs ont leur mot à dire, car ils pourraient en être les premiers bénéficiaires.

Il s'agit donc d'un agenda très chargé pour une tâche qui n'est pas simple. L'expérimentation pourrait être la solution dans

7. Scalable Vector Graphics, voir <http://www.w3.org/Graphics/SVG/>

bien des cas; pour y arriver la meilleure méthode consiste probablement en la création d'une nouvelle revue savante purement électronique, dans une discipline qui se prête bien à la diffusion de différents types d'information.

Conclusion sur les modèles techniques de publication

Cette première partie avait pour objectif de présenter les différentes techniques reliées à l'édition électronique. Nous avons surtout cherché à montrer que sans un modèle technologique solide et orienté vers le document électronique et la réutilisation de l'information, il est impossible d'exploiter un tant soit peu les possibilités des documents électroniques. Les aspects importants de ce modèle sont les méthodes de production et les formats de documents (pour la gestion, la diffusion et la conservation), et ils doivent mener à une exploitation optimale des possibilités de l'électronique. Heureusement, un tel modèle existe et il a été expérimenté dans le cadre du projet Érudit. Il s'agit de baser la chaîne de traitement sur un document structuré, en format XML, à partir duquel les différents formats de diffusion, y compris sur support imprimé, sont produits (d'autres éditeurs sont aussi arrivés à cette conclusion, voir entre autres Kasdorf 1998). Les détails techniques de ce modèle seront présentés dans la seconde partie.

L'existence d'un tel modèle ne signifie malheureusement pas l'absence de tout problème ou la réalisation sans douleur de projets d'édition électronique. Les véritables documents électroniques, qui incluent de l'information dynamique et interactive, sont des objets avec lesquels on doit continuer à se familiariser afin de trouver des applications et des formats adéquats. Mais, surtout, il est important de revoir notre conception de ce qu'est un document et de le considérer plutôt comme une *interface à un réservoir d'information*. D'une part, l'interface peut être individualisée pour chaque utilisateur, à chaque consultation, d'autre part, le réservoir d'information peut être en constante modification. Cela demande un changement de culture important chez les éditeurs et les auteurs et, à un degré moindre, chez les lecteurs-utilisateurs. Ce changement s'exprime, bien entendu, par de nouvelles chaînes de traitement ou de nouveaux modèles techniques, mais il doit également s'exprimer par de nouvelles mentalités.

Présentation d'une chaîne de traitement : l'exemple du projet Érudit

L'histoire nous a montré que tout nouveau média a débuté en tentant d'abord de calquer un moyen de diffusion existant. Éventuellement, ce nouveau média évolue et arrive à se définir et à développer ses propres caractéristiques devenant ainsi réellement novateur. Ainsi, pour la plupart des initiatives et des innovations dans le domaine de la publication scientifique, le modèle de la revue savante sur le Web commence par une transposition du modèle imprimé déjà existant.

Pour en assurer la réussite, la « transition » du format papier à l'électronique implique, d'une part, de travailler selon des façons de faire déjà existantes et, d'autre part, de graduellement en implanter de nouvelles, propres aux impératifs et aux possibilités d'un traitement électronique. Dans le cadre du projet pilote Érudit, cette nouvelle chaîne de production a été réalisée parallèlement au processus de production courant des revues. Il a été établi qu'on tenterait de reproduire le plus fidèlement possible la signature (apparence) et la structure (construction) qu'affichait la forme imprimée des revues participantes, afin d'assurer une transition en douceur de même que pour respecter l'intégrité des contenus. D'un point de vue technique, cela s'apparente à faire un traitement rétrospectif des documents puisque aucune intervention de notre part ne pouvait être faite préalablement ou parallèlement à la création du document. Ainsi, la méthodologie mise en place et exposée ici est-elle le fruit d'un mélange de conditions, de contraintes et de possibilités dictées par cette situation. Cette méthodologie, on le comprendra aisément, est bien différente de celle qui pourrait être mise en place pour la production d'une revue uniquement électronique, et cela serait d'autant plus vrai s'il s'agissait d'une toute nouvelle revue. Dans ce dernier cas, toutes les conventions et interventions nécessaires à la production d'une revue électronique peuvent être considérées et implantées dès le début de sa conception.

Nous devons aussi souligner que cette chaîne est basée sur la norme SGML, mais en fait elle pourrait très bien l'être sur XML, la norme plus récente. En effet, aucune fonctionnalité de SGML non présente dans XML n'a été utilisée. Par conséquent, dans la description qui suit, les expressions « SGML » et « XML » sont interchangeables.

Définition du type de document (DTD)

Au début d'un projet basé sur SGML, le choix ou la création d'une *Définition du type de document* (DTD) est une étape primordiale. La DTD est le fondement d'une application SGML. C'est elle qui détermine de quelle façon les documents seront représentés, les traitements qui seront possibles, etc. En quelque sorte, il s'agit du véritable « format » des documents. Les réponses aux questions suivantes orienteront la création ou le choix d'une DTD :

- Quels sont les types de documents à traiter ? Quels sont les types de documents semblables ?
- Quelles sont les composantes structurelles des documents ? Quels sont les autres types d'éléments logiques apparaissant dans chaque type de document ?
- En plus des contenus textuels, quelles autres informations ou propriétés peuvent être assignées à chaque type d'élément ?
- Quelles sont les relations logiques entre chacun des éléments ?
- Que veut-on faire de l'information ? Quelles sont les types de structures et de relations que l'on veut encoder dans le balisage SGML de façon à pouvoir répondre aux besoins d'échange (partage), de repérage, de diffusion et de réutilisation de l'information ?

Cet exercice en est un d'analyse de besoins en fonction des coûts encourus et des bénéfices retirés. Ainsi un balisage fin et hautement structuré, insufflant de ce fait une « intelligence » aux documents, permettra une exploitation plus performante des données tandis qu'un balisage plus grossier, effectué à moindre coût, trouvera une possibilité de réutilisation réduite ou demandant certaines interventions.

En matière de revues savantes, l'équipe des publications électroniques

des Presses de l'Université de Montréal a eu, dans des projets antérieurs à celui-ci, l'opportunité d'étudier la question du choix de la DTD pour ce type de publication[8]. La DTD ISO 12083 Article ayant été décrétée trop limitée, on a jugé souhaitable d'adapter cette dernière aux besoins des revues savantes. Du noyau de la DTD ISO 12083, auquel on a greffé des fragments d'autres DTD normalisées — CALS (*Continuous Acquisition and Life Cycle Support*), TEI (*Text Encoding Initiative*), DTD de la revue *Earth Interactions*[9] —, nous en sommes arrivés à créer la DTD des Presses de l'Université de Montréal (alias DTD PUM), laquelle devait décrire la structure d'un article de périodique d'une revue savante. On a également procédé à certains ajouts maison, concernant entre autres la gestion des différentes langues, le découpage des références bibliographiques, etc.

Description des étapes de la chaîne de traitement

Analyse préliminaire

Avant d'utiliser la chaîne de traitement proprement dite, un examen attentif de plusieurs numéros déjà parus doit être mené pour chaque revue. Cet examen doit être fait pour chaque nouvelle revue. Idéalement, l'analyse doit porter sur une période rétrospective suffisamment longue et un nombre de numéros suffisamment important pour englober l'ensemble des caractéristiques et variations de la revue. Cette analyse servira à identifier les types de documents (articles, comptes rendus, notes, études, etc.) et leur structure sémantique[10] MHV (titres, résumés, sections, subdivisions, tableaux, équations, illustrations, citations, renvois, références, notes infrapaginales, etc.).

C'est aussi l'occasion d'observer la régularité et la conformité de l'application du protocole de rédaction en vigueur, ce qui a un impact important sur la qualité. En principe, si les documents présentent une structure uniforme et constante (à l'intérieur d'un même document et d'un document à l'autre), si le protocole de rédaction a dûment été respecté et si l'étape du choix ou de la création de la DTD a été bien faite, les composants sémantiques rencontrés, leur nombre et l'ordre dans lequel ils se présentent devraient tous être conformes à la DTD. Dans la situation où la correspondance entre la DTD et certains éléments logiques se retrouvant dans les documents est impossible, il faudra apporter des modifications soit à la DTD, soit aux documents, soit à la fois à la DTD et aux documents. Cette intervention mérite une réflexion importante puisqu'elle est lourde de conséquences. Dans un contexte de conversion rétrospective, il est impossible d'intervenir sur le contenu des documents. Les changements à la DTD doivent être faits en considérant un éventail de besoins et non seulement le besoin précis qui nous amène à vouloir modifier la DTD. Aussi, les changements à la DTD doivent être additifs et non correctifs, de façon à garder une certaine compatibilité avec les documents produits antérieurement à ces changements.

Outre les modifications de nature sémantique, les amendements doivent s'inscrire plus largement dans l'utilisation souhaitée, tant à court qu'à plus long terme, des documents. En effet, la mise en page en fonction de différents médias (papier, écran, etc.), la recherche structurée, la réutilisation de certaines parties de documents, etc. dépendront toutes d'une indication SGML (un élément spécifique, une valeur d'attribut, etc.). En principe, cette réflexion a dûment été faite lors de l'étape du choix ou encore de la conception de la DTD de sorte que les éléments logiques devraient tous s'harmoniser avec la DTD. Cependant, puisque aucun contrôle n'a pu être exercé sur la structuration initiale des documents comme il a été expliqué plus haut, il se peut que certains éléments logiques ne trouvent pas de correspondance dans la DTD.

Il faut également discerner les caprices de la mise en page papier actuelle, faite de façon manuelle la plupart du temps, des véritables besoins de discernement visuel des composants logiques. Ainsi, une information SGML doit être prévue pour faire en sorte qu'une liste ordonnée soit rendue de façon différente d'une liste à puces. On pourra ainsi, à l'étape de production d'une sortie papier, produire respectivement un numéro séquentiel ou un symbole tel un rond plein (« • ») devant chaque item de liste.

En théorie, la production du SGML et les ajustements à la DTD ne devraient pas être faits en fonction des limites des outils employés ; cependant, en pratique, on ne peut ignorer cet aspect. Par exemple, une formule mathématique peut être finement balisée selon un modèle de DTD approprié (fonction, numérateur, dénominateur, arguments, etc.). En pratique cependant, plusieurs logiciels de notre chaîne de traitement ne peuvent reconnaître et traiter adéquatement une formule reproduite sous cette forme. On se rabattra alors sur une représentation graphique de la formule, c'est-à-dire l'insertion d'un fichier image. Ceci peut demander des modifications à la DTD. Autre exemple : l'emploi de jeux de caractères étrangers, bien qu'ils puissent être représentés en SGML (entre autres, par l'emploi d'entités caractères codées en Unicode[11]), peut nous faire rencontrer les limites des outils de traitement. Ici, encore, on préconisera une solution de rechange impliquant des modifications au SGML produit et, conséquemment, à la DTD. Si le nombre de limitations rencontrées par les outils est très important, on décidera, afin de ne pas « corrompre » abusivement la source SGML, de produire (automatiquement, de préférence) des versions SGML transitoires.

Nous avons recours à une version SGML transitoire pour la production du PDF (une étape plus en aval de la chaîne qui sera expliquée plus loin). Cette version transitoire sert à réordonner et qualifier certains éléments de façon à accommoder certaines caractéristiques propres aux revues (par exemple, le fait de présenter les résumés à la toute fin de l'article plutôt qu'au début). Elle est produite automatiquement avec le logiciel *OmniMark*.

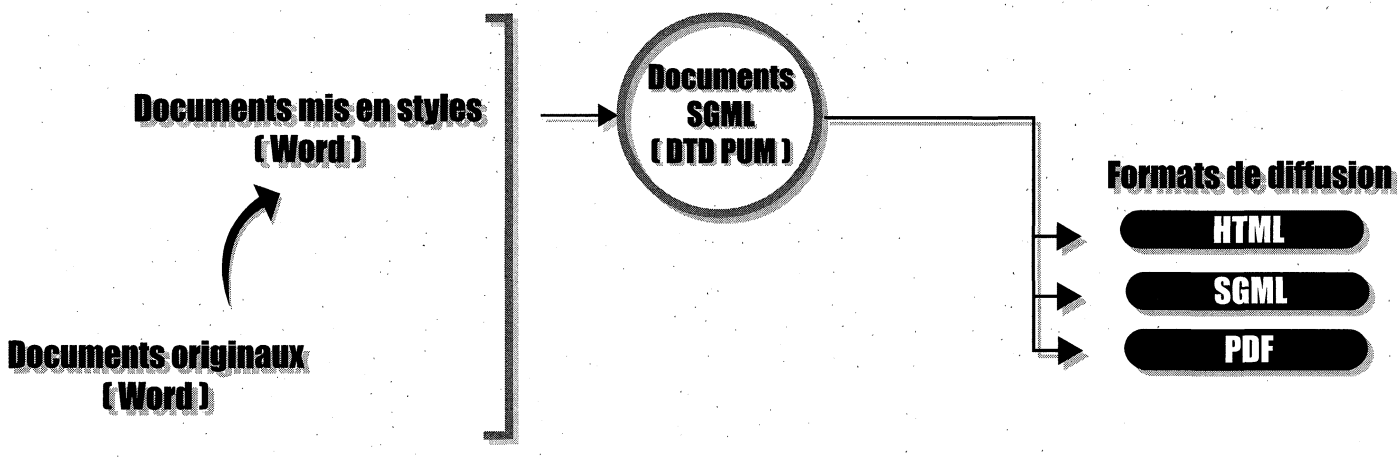
Prétraitement

Mise en styles et préparation des textes

La chaîne de traitement proposée comporte une étape de prétraitement des documents originaux (Figure 1). Cette

8. Voir *Un nouveau modèle de publication électronique* ; <URL :http ://www.pum.umontreal.ca/publ_electr/vision.html>
9. <URL :http ://EarthInteractions.org>
10. Nous employons dans cet article l'expression « structure sémantique » pour souligner l'organisation d'un document textuel basé sur le sens de ses composants par opposition à leur apparence.
11. La norme Unicode est un jeu de caractères étalé sur 16 bits, ce qui permet la représentation d'un maximum de 65 536 caractères différents. Elle correspond au premier plan à 2 octets (le BMP, Basic Multilingual Plane), d'une norme plus universelle (basée sur 4 octets) soit la norme ISO 10646. Voir <URL :http ://www.unicode.org>.

Figure 1 : Chaîne de traitement du projet Érudit



étape consiste à appliquer un ensemble de styles, contenus dans un modèle de document (feuille de style), à un document original en format *MS Word*. L'utilisation de styles sur des parties de texte inscrit, dans le format intrinsèque du traitement de texte, des codes qui nous permettront d'automatiser le balisage en SGML.

C'est à cette étape que l'on insuffle une certaine « intelligence » au document. Ainsi, le modèle de document et les styles qu'il contient ne sont pas destinés à modifier l'apparence des documents, comme l'ont d'abord pensé les concepteurs de cette fonctionnalité, mais plutôt à identifier certaines parties des documents. Par exemple, nous chercherons à identifier les titres, les auteurs, les références bibliographiques, les résumés, etc. Il est important de comprendre que nous n'utiliserons pas les styles à des fins de mise en page, mais bien d'identification des différents composants du texte.

Outre l'application des styles, l'uniformisation et la normalisation des textes doivent être effectuées. Ainsi, les tableaux, les listes et les notes infrapaginales doivent avoir été produits avec les fonctionnalités appropriées du traitement de texte (en d'autres mots, que ce soient de vrais tableaux et non des chaînes de texte séparées par des tabulations, par exemple). Nous analysons, à l'étape de la programmation, l'ensemble des codes du traitement de texte : les liens entre les appels de notes et les notes, la largeur relative des colonnes des tableaux, l'alignement des contenus des cellules des tableaux, etc., et nous reproduisons cette information dans les documents SGML. Au besoin, certains composants doivent être réordonnés (par exemple, si l'auteur a ajouté son résumé à la fin, on devra le repiquer au début du document, à la suite des éléments de titres et d'auteurs, tel que le prévoit la DTD).

Nous avons développé un modèle de document pour chacune des revues. Ces modèles contiennent en moyenne de 30 à 50 styles différents en fonction du nombre de composants retrouvés. Parmi ces styles, certains sont de type *paragraphe* (blocs de texte) et d'autres de type *caractère* (chaînes de texte à l'intérieur des blocs).

L'étape de mise en styles pourra être réalisée par la personne responsable de la préparation des textes à publier, soit habituellement une personne faisant partie du comité de production de la revue. À cet effet, nous avons, dans le cadre du projet Érudit, entrepris de donner des séances de formation personnalisée aux responsables des revues, lesquelles portaient sur la préparation des documents et l'application des feuilles de style. Notre expérience démontre que cette étape est plus facilement réalisable par des personnes maîtrisant bien leur outil de traitement de texte. Un certain contrôle de qualité doit être effectué sur les premiers documents stylés.

Une fois stylé, le document servira d'intrant (*input*) à l'étape de conversion SGML. C'est pourquoi il doit s'agir d'une version finale du texte, c'est-à-dire qu'il doit être complet et déjà comprendre toutes les corrections nécessaires. À cet effet, nous avons rencontré quelques problèmes puisque le traitement traditionnel des revues comporte une ou plusieurs étapes de correction d'épreuves une fois le document mis en page à l'aide d'un logiciel de PAO (tel *PageMaker* ou *XPress*). Les corrections sont alors effectuées directement dans ces logiciels et ne sont pas repiquées dans la copie de traitement de texte initiale. Notre système de traitement préconise une utilisation de fonctionnalités appropriées du traitement de texte (styles, notes, tableaux, etc.). De plus, les corrections de nature textuelle (orthographe et syntaxe, de même que la plupart des corrections de nature typographique) doivent être faites dans le fichier de traitement de texte. Il faut prendre en considération le fait que ce fichier sert à la production de documents d'archivage, de documents dérivés (sommaires, tables des matières, listes de résultats de recherche, etc.) ainsi que de plusieurs formats de diffusion. Ce fichier source, de par son contenu et sa structure, doit donc être le plus achevé possible. Cela exige de la personne qui prépare les textes, de bonnes connaissances linguistiques et typographiques. Les erreurs révélées en aval dans la chaîne de production doivent être corrigées dans les fichiers de traitement de texte.

Épreuves électroniques

Depuis la fin du projet Érudit, nous avons testé l'intégration du processus de révision des épreuves dans la chaîne de traitement. Dans un processus plus traditionnel, cette révision se fait habituellement sur les épreuves imprimées des articles, une fois la mise en page effectuée à l'aide de logiciels de PAO. Les corrections apportées peuvent être à la fois de nature orthographique, syntaxique, typographique et peuvent aussi concerner des aspects du montage. La diffusion du même contenu sur plusieurs supports de diffusion nous amène à discerner entre les corrections devant être repiquées dans le fichier initial de traitement de texte (de façon à ce que l'ensemble des formats de diffusion profitent de ces modifications), des corrections s'appliquant uniquement à la mise en page papier. Nous produisons donc maintenant des « épreuves électroniques », c'est-à-dire des sorties imprimées obtenues à partir d'un logiciel de PAO, soit *FrameMaker+SGML* (voir plus loin pour la description de cette étape) sans avoir à effectuer d'importantes interventions manuelles au niveau du montage. Sur ces épreuves sont apportées les corrections se rapportant au texte (et non à sa mise en page en vue de l'imprimé), lesquelles seront ensuite repiquées dans le fichier *Word* initial qui servira à la régénération des différents formats produits. Dans la perspective où plusieurs formats différents (SGML, HTML, PDF) sont produits à partir d'un même fichier initial, on comprendra l'intérêt d'effectuer un maximum de corrections en amont (donc dès la révision du fichier *Word*) et non en aval, sur un format en particulier plutôt que sur un autre.

Traitement des images

Les illustrations que peuvent comporter les articles parviennent habituellement aux responsables des revues sous différentes formes : photos, impressions laser, velox, etc. et plus rarement sous forme électronique : fichiers vectoriel ou bitmap provenant de logiciels graphiques et images numérisées. Traditionnellement, ces images sont insérées dans le texte au moment du montage de la copie à imprimer par le typographe qui les numérise et les retouche, ou encore par le technicien de l'atelier de préimpression qui les reproduit par procédé photographique.

L'intégration des images dans la chaîne de traitement Érudit est de beaucoup facilitée si ces documents nous parviennent en format numérique. Les images sont traitées comme des fichiers séparés qui sont référés à partir des documents.

Les secrétariats des revues n'étant habituellement pas équipés en matériel et savoir-faire pour manipuler des images, ce type d'intervention sera effectué par le fournisseur de services, en l'occurrence l'équipe d'Érudit. Cependant, les comités de revues devraient tenter d'obtenir des auteurs, autant que faire se peut, une version des images en format électronique. Ces images devraient toujours être de très bonne qualité (avec une haute résolution et, si désiré, en couleurs) afin de servir à l'archivage. C'est à partir des images de bonne qualité qu'on effectuera le traitement (surtout des modifications de formats de fichiers, des baisses de résolution, des sauvegardes en noir et blanc) en vue des différents modes de diffusion (archivage, Web et imprimé). Afin de standardiser le processus de traitement des images, nous avons établi des procédures normalisées. Enfin, bien que cela ne se soit pas encore produit, un type d'intervention de même nature est à préconiser pour d'autres formats non textuels (vidéo, son, images 3D, etc.).

Nous avons exposé dans cette section les différents éléments portant sur le prétraitement des textes, soit l'application des styles et le traitement des images. Une fois le document prétraité, il est prêt à être converti en SGML.

Conversion en SGML

Puisque la feuille de style utilisée pour styler les documents a été conçue en fonction d'un passage efficace vers un document SGML respectant la DTD PUM, une conversion relativement automatique est possible. Cette conversion, qui consiste en l'interprétation des informations existantes et l'ajout d'informations dérivées, requiert cependant un outil spécialisé. Il serait, à ce point-ci de la discussion, important d'expliquer la différence entre une conversion de données au sens SGML et une traduction de données.

Une traduction est une opération qui consiste à prendre un ensemble de données (information et mise en forme confondus) d'un format propriétaire et le traduire en un ensemble équivalent qui peut être interprété et édité dans un autre logiciel utilisant un format propriétaire (par exemple, passer un document de *Word* à *WordPerfect*). Au contraire, une conversion vers SGML implique d'établir un lien entre un document dans un format propriétaire (où la structure logique est habituellement perçue de façon visuelle par le lecteur) à un document SGML « intelligent » (où la structure logique est codée de façon explicite, suivant une DTD donnée).

Une conversion enrichissante (traduction libre de *up conversion*) est une conversion d'un format plat vers un format sémantique structuré (par exemple, d'un format *WordPerfect* vers un format SGML) ou, plus précisément, une conversion de données textuelles d'un format d'encodage arbitraire vers une instance SGML valide. Une conversion appauvrissante (traduction libre de *down conversion*), quant à elle, consiste en une conversion d'un format logique vers un format propriétaire (par exemple, d'un format SGML vers un format *MS Word*). Ce type de conversion est la clé du succès de SGML puisqu'elle assure une complète réutilisation des données. À partir d'un format SGML, la conversion peut être facilement automatisée tout en conservant l'intégrité des données.

Il existe plusieurs produits logiciels aptes à convertir des textes d'un format à un autre. Ces produits sont disponibles essentiellement sous la forme de programmes ou de solutions paramétrables[12]. Nous avons opté pour le langage de programmation spécialisé OmniMark®.

Nous effectuons la conversion en deux étapes, dont la première est réalisée à l'aide d' un programme en langage OmniMark disponible gratuitement[13]. La seconde étape se charge d'ajouter une structure sémantique normalisée pour la description d'articles savants et introduira de l'information supplémentaire à valeur ajoutée. Ainsi nous créons, sans aucune intervention manuelle, des liens entre les appels de références dans le texte et les références bibliographiques à la fin du

12. Pour une liste exhaustive de ce genre de produit, consulter la rubrique « Conversion Program » dans *Survey of software for structured text* par Eila Kuikka et Erja Nikunen ; <URL :http ://www.cs.uku.fi/~kuikka/systems.html>
13. Il s'agit de *RTF2XML* (auparavant *RTF2SGML*), développé par Rick Geimer. <URL :http ://www.xmeta.com/omlette/>

texte. Pour ce faire, il faut analyser méticuleusement les syntaxes de références utilisées par chaque revue [par exemple « (Vaugeois et Hubert 1993a) » ou « (VaHu93a) »], afin d'établir des patrons de reconnaissance sur lesquels on peut ensuite bâtir des règles de programmation. Cette deuxième étape est effectuée par un programme en langage OmniMark réalisé dans le cadre du projet Érudit.

Formats de diffusion

Les formats de diffusion choisis doivent permettre de satisfaire les habitudes de lecture et les besoins de consultation qui diffèrent d'un usager à l'autre. Ainsi, nous préconisons la diffusion des documents en plusieurs formats. Notre modèle de diffusion est basé sur le document dans sa version SGML. Ce document électronique unique constitue la source de plusieurs sous-produits qui sont autant de formats de sortie possibles. Les supports de diffusion qui sont compatibles avec ces formats sont divers : diffusion en ligne (Web), diffusion sur cédérom, diffusion sur papier. Dans le cadre du projet Érudit, nous avons considéré trois formats de sortie (HTML, SGML et PDF). Nous avons également décidé de rendre disponible le format SGML, ce dernier se prêtant très bien à la consultation dans la mesure où les navigateurs SGML sont capables d'exploiter la structure des documents de façon à offrir un environnement de lecture plus sophistiqué et complet. Enfin, le format PDF permet la diffusion de documents électroniques reproduisant la mise en page choisie par les diffuseurs. Ce format est tout à fait approprié pour l'impression à distance. Nous envisageons à court terme d'offrir également le format XML qui peut être consulté à l'aide des versions 5 des navigateurs *Internet Explorer* et *Netscape Navigator*.

Un principe de base de la philosophie SGML voulant que l'information, une fois balisée en SGML, puisse être réutilisée plusieurs fois et cela avec différentes saveurs, tant dans l'intégralité du contenu (en tout ou en parties) que dans la forme du contenant (dans un format particulier), il est tout naturel de convertir les fichiers SGML en fichiers HTML. Il s'agit en fait d'une conversion SGML (DTD PUM, soit la DTD développée aux Presses de l'Université de Montréal pour répondre aux besoins de balisage des revues savantes) vers SGML (DTD HTML). Ainsi, les articles sont convertis automatiquement en HTML à l'aide d'un autre programme OmniMark. Nous produisons, de la même façon, une table des matières pour chaque numéro de revue. Cette table des matières ne provient pas d'un document initial unique saisi sous forme de table des matières, mais plutôt de la concaténation d'informations récupérées dans l'ensemble des articles (titres, sous-titres, auteurs) en format SGML. C'est un bon exemple de réutilisation de l'information.

Puisque l'ensemble des documents produits sont basés sur la DTD PUM, la conversion du SGML vers le HTML est réalisée par un seul programme de conversion pour l'ensemble des revues. Ce programme comporte cependant quelques particularités de présentation propres à chaque revue, et d'une telle pratique résulte une apparence assez semblable pour chacune des revues. La fabrication d'une signature, d'une maquette différente pour chaque revue est réalisée à l'aide d'une feuille de style CSS (*Cascading Style Sheet*), fonction de plus en plus supportée par les navigateurs Web.

Pour l'instant, nous stockons sur notre serveur Web la version HTML préconvertie. Cependant, la conversion, tant pour les articles que pour la table des matières, pourrait aisément être faite « à la volée » (*on the fly*) puisqu'il n'y a aucune intervention manuelle sur les fichiers HTML. On pourrait facilement faire en sorte qu'au moment où l'utilisateur, par le biais de son client Web, fait la demande d'un fichier HTML (par exemple, un article donné), le serveur (sur lequel serait installé OmniMark) convertisse le fichier SGML en fichier HTML. Ceci aurait pour avantage d'occuper moins d'espace disque (environ la moitié de moins, puisque seule la version SGML serait stockée et que le rapport SGML/HTML doit plus ou moins être égal à 1). Il resterait à s'assurer de délais de réponses acceptables.

SGML

Les utilisateurs peuvent visionner les documents SGML à l'aide de *Panorama Viewer* de Interleaf[14], un navigateur SGML qui s'installe comme module externe (*plug-in*) d'un navigateur Web, tel *Netscape Navigator*.

Panorama Viewer permet de créer des liens entre différentes parties des documents ou différents documents. Il permet également de présenter des tables des matières dynamiques qui constituent des aides à la navigation. Une autre fonction de *Panorama* offre la possibilité d'annoter n'importe quelle partie d'un document, y compris une zone d'une figure déterminée par le lecteur. Enfin, il permet d'ajouter des signets dans un document.

La diffusion des articles en format SGML requiert peu d'efforts puisqu'il s'agit du format natif des documents. Pour être plus exact, nous diffusons une version SGML à peine différente de la version produite lors de l'effort de conversion enrichie décrite précédemment. Cette différence minime tient exclusivement à la non-reconnaissance, par le navigateur SGML, de certains caractères étrangers et symboles particuliers. *Panorama Viewer* étant en mesure de lire du SGML, il faut simplement lui indiquer la mise en forme propre aux différents éléments rencontrés. Ces instructions de formatage sont données sous la forme d'une feuille de style. La création d'une feuille de style ne s'effectue normalement qu'une seule fois par revue (voire même par DTD utilisée si on désire des produits normalisés et identiques). Ainsi nous avons créé des feuilles de style et des formats de tables des matières dynamiques (*navigators*) pour chaque revue, de manière à donner à chacune une signature caractéristique.

PDF

L'intérêt du PDF réside dans l'obtention d'un document électronique arborant un format de présentation semblable à l'imprimé. Un tel format permet aux lecteurs d'imprimer une version « mise en page papier » des articles diffusés en ligne. Les navigateurs SGML ou HTML, dont nous avons discuté précédemment, permettent également l'impression, mais la qualité de la mise en page n'est pas aussi complexe et soignée puisque les feuilles de style des navigateurs Web sont destinées avant tout à la présentation à l'écran. Notons cependant qu'il existe des feuilles de style (CSS et XSL) normalisées où on a prévu une mise en page sophistiquée en fonction de différents médias (écran, terminal réduit, imprimé, synthèse vocale). À moyen terme, on pourrait entrevoir que les

14. <URL :http ://www.interleaf.com/Panorama/page3.html>

formats HTML et PDF seront remplacés par un seul document source XML couplé à des feuilles de style qui seront fonction du média employé.

« La grande majorité des logiciels de PAO sur le marché (les plus connus étant *PageMaker* et *Xpress*) permettent une sauvegarde en format PDF pour diffusion sur le Web. Ces logiciels ne sont cependant pas en mesure de lire en intrant (*input*) un fichier SGML ni de l'interpréter correctement. Il existe toutefois des outils performants capables à la fois de prendre en intrant du SGML et de permettre une mise en page professionnelle (*FrameMaker+SGML*[15] de Adobe, *Interleaf 6 SGML*[16] de la compagnie Interleaf, et *ADEPT Publisher*[17] réalisé par ArborText). Il s'agit toutefois de produits dispendieux et assez complexes à manipuler.

Notre choix s'est arrêté sur le logiciel *FrameMaker+SGML* en raison de son coût abordable et de sa plus grande facilité de paramétrage. *FrameMaker+SGML* permet la création, la modification et la publication de documents SGML dans un environnement convivial. Dans le cadre du projet Érudit, nous utilisons uniquement les fonctionnalités d'importation du SGML et de mise en page sophistiquées offertes par l'outil. Il s'agit, ici aussi, de définir un mécanisme d'application de style pour chaque élément en fonction de son contexte SGML. Cette feuille de style est dirigée vers la mise en page sur papier (et non à l'écran) et, à cet effet, *FrameMaker+SGML* offre la possibilité de paramétrer plusieurs des caractéristiques requises par l'imprimé (par exemple, la gestion des veuves et orphelins, des paragraphes solidaires, les colonnes multiples, les notes en bas de page, les patrons de césure appropriés à la fin des lignes, les formes des en-têtes et bas de pages, etc.).

Chaque revue possède une feuille de style différente, laquelle a été conçue en se basant sur la maquette de la forme imprimée déjà existante des revues participantes. Les fonctionnalités de *FrameMaker+SGML* nous permettent sans difficulté de reproduire de façon presque identique l'aspect habituel des revues. L'application d'une feuille de style exhaustive et bien conçue (qu'on arrive généralement à obtenir après un processus de réajustement lors des premières productions) permet d'automatiser à 90 % le montage du document, ce qui constitue un excellent rendement. On doit toutefois faire certaines interventions manuellement afin de compléter le travail (par exemple, déplacer des figures, condenser l'espacement du texte ça et là pour forcer un saut de page, fractionner des notes de bas de page trop longues, etc).

Une fois le document monté avec *FrameMaker+SGML*, il ne reste qu'à produire une version PDF à l'aide du logiciel *Acrobat* de Adobe. Il s'agit là d'une opération sans grande difficulté qu'on peut apparenter à un simple « sauvegarder sous PDF ». On peut ainsi produire un premier fichier PDF de faible résolution, destiné à être visionné sur le Web, et un second fichier PDF de meilleure résolution (ou encore un fichier PostScript), pouvant être acheminé à l'atelier de prépresse en vue de l'impression des exemplaires papier.

Pour les usagers, la consultation des documents PDF se fait avec le logiciel *Acrobat Reader*, qui existe comme module externe (*plug-in*) pour les navigateurs Web. On peut consulter les documents PDF à l'écran avec ce logiciel, mais ils sont plutôt destinés à être imprimés localement par les usagers.

Dans le cadre du projet Érudit, nous avons produit uniquement des documents PDF destinés au Web, à raison d'un fichier unique par article (et non d'un fichier regroupant tout un numéro d'une revue), puisque les revues poursuivaient pendant la même période, le traitement habituel pour la production de la version imprimée de la revue.

Diffusion sur le Web

Pour diffuser des documents sur le Web, nous avons créé un site Web[18], hébergé sur notre serveur Web. Ce site est la porte d'entrée principale pour l'accès aux articles.

La diffusion en ligne (et également sur cédérom) permet, pour le lecteur, d'accéder aux textes à l'aide d'un outil de recherche. Pour ce faire, on doit implanter un outil de recherche en texte intégral qui peut indexer des documents SGML, tout en conservant l'information sur la structure, de façon à permettre aux usagers d'exprimer des contraintes sur la structure dans leurs requêtes. On peut, par exemple, préciser la recherche de textes comportant le nom d'un chercheur, mais uniquement lorsque celui-ci se retrouve dans une bibliographie et non en tant qu'auteur de l'article. Dans le cadre du projet Érudit, nous avons installé, sur le site Web, un prototype de moteur de recherche qui doit encore être amélioré. La collection indexée comprend l'ensemble des articles ayant été publiés par les revues pendant la durée du projet pilote. Il s'agit d'un mode d'accès au texte très performant que seule la version électronique permet d'obtenir.

Recommandations

Dans la mesure où notre mandat consistait moins en la mise au point d'une solution théoriquement optimale qu'en la conception d'une chaîne de traitement adaptée aux besoins du projet (reproduction de la forme papier déjà existante, impossibilité d'intervenir en amont de la chaîne de production, optimisation du traitement par l'adoption d'une DTD unique, production de divers formats de diffusion), nous avons opté pour une solution comprenant plusieurs sorties intermédiaires, et cela, de façon à atteindre de bons résultats en matière d'automatisation et de qualité des textes produits. Nous estimons avoir atteint à 95 % ces objectifs de production.

Il va sans dire que la DTD a dû être constamment remaniée afin de s'adapter à l'ensemble des documents à produire. Nous avons dû opter pour une DTD qui, au fur et à mesure de l'avancement du projet, devenait plus permissive, de manière à générer, en bout de ligne, du SGML valide, respectant l'intégrité et la présentation des contenus des articles.

Le projet Érudit aura mis en lumière l'extrême importance, dans une perspective d'automatisation des opérations de la chaîne documentaire, de définir une structure, une présentation et, dans une certaine mesure, un contenu normalisé des documents à traiter. La production même de textes destinés à être balisés en SGML peut donc faire l'objet de certaines recommandations, dont plusieurs s'avèrent applicables dans une optique de sensibilisation des comités de revues et de formation du personnel chargé de la révision et de la correction des textes. C'est ainsi qu'il y aurait lieu de définir (ou renforcer), pour chaque revue, un modèle normalisé quant

15. <URL :http ://www.adobe.com/prodindex/frame-maker/prodinfosgml.html>
16. <URL :http ://www.interleaf.com/products/p_sgml.html>
17. <URL :http ://www.arbortext.com/Products/ADEPT_Series/adept_series.html>
18. <URL :http ://www.erudit.org>

à la présentation des textes et surtout de l'appliquer de façon rigoureuse.

Cet aspect, d'une importance capitale, pourrait s'effectuer notamment par l'élaboration de documents du type « guide ou protocole de rédaction » et l'encouragement des auteurs à utiliser une feuille de style calquée sur la DTD, laquelle serait rendue disponible par la revue sur son site Web ou encore directement envoyée aux auteurs par courriel. Mentionnons également l'importance d'inciter fortement, voire d'exiger des auteurs, des versions électroniques de bonne qualité des illustrations accompagnant les articles. Ces versions électroniques existent souvent, mais pour plusieurs raisons, seules les versions imprimées parviennent aux comités des revues.

Nous avons entièrement basé notre chaîne de production sur le fait que les documents initiaux sont en format de traitement de texte, en l'occurrence MS Word, parce qu'il s'agit de l'environnement de saisie pour la majorité des auteurs. Cependant, nous avons constaté qu'un nombre très important d'interventions est fait sur le texte au cours du processus éditorial, de sorte que la somme des modifications, des corrections et des ajouts équivaut à la ressaisie d'une grande partie du texte. Puisque de toute façon, on devra allouer beaucoup de temps et d'efforts au remaniement du texte, on pourrait également envisager, dès le début de la chaîne de traitement, qu'un responsable de l'édition ouvre, puis travaille le document dans un éditeur SGML et produise ainsi une version SGML des textes. Les avantages/caractéristiques d'un éditeur SGML/XML sont nombreux: validation de la syntaxe SGML/XML, listes d'autorités et alias pour des valeurs d'attributs et de contenus d'éléments, environnement personnalisé de création, direction dans la structuration du texte en indiquant quels éléments sont permis et à quelles places dans l'avancement de la rédaction, etc. Ceci permet à l'auteur de gagner du temps et obtenir des résultats plus homogènes (tant intra que interdocumentaires). Le produit obtenu est directement du SGML/XML et n'a pas besoin de conversions supplémentaires. Le travail de diffusion à investir en aval de la production du document est donc considérablement réduit.

Il s'agit de l'une des adaptations que l'on pourrait faire pour que cette chaîne de traitement permette aux revues d'exploiter de plus en plus les possibilités des documents électroniques. Pour l'instant, aucune des revues traitées dans le cadre de ce projet n'a véritablement utilisé des caractéristiques de l'information que l'on ne peut reproduire sur papier, mais nous savons que lorsque le temps viendra, notre modèle pourra toujours fonctionner.

Vers un produit électronique complet

Tel que mentionné au tout début de cet article, la chaîne de traitement mise en place par l'équipe d'Érudit s'apparente plus à une transposition, plutôt qu'à une transition, du processus de production de l'imprimé à la production de la forme électronique. De nombreux efforts ont dû être investis, auprès des comités des revues, dans la formation sur l'utilisation des fonctionnalités de traitement de texte ainsi que dans la standardisation des fichiers soumis. Nous sommes persuadés que nous devons aller de l'avant vers la création de véritables produits électroniques qui, tout en respectant les principes de base de la communication savante, tirent tous les avantages de l'hypermédia et des réseaux. On pense à l'ubiquité des publications, à la recherche plein texte (recherche structurée), au court délai de parution des articles, à la diffusion d'articles en devenir (articles en versions intermédiaires, *work in progress*), à la diffusion sélective de l'information (DSI), au multimédia, à l'ajout de matériel supplémentaire tel données brutes, images couleurs, vidéo, etc., à l'inclusion de « données actives » telles équations et modèles de simulation pouvant être manipulées par l'utilisateur, à la mise en place de forums d'échanges, à la publication interactive (soit le concept de *Scholarly Skywriting,* développé par Harnad, 1990) qui permet l'ajout de commentaires par les pairs et fait de l'article un séminaire permanent, au monitoring des utilisations par les usagers, etc. Certaines de ces valeurs ajoutées sont déjà implantées, d'autres devront être optimisées, d'autres encore relèvent d'un avenir plus ou moins lointain. Afin d'atteindre cet objectif de produit électronique complet, nous croyons que nous devons sensibiliser les responsables des revues ainsi que, de façon indirecte, les auteurs, aux multiples possibilités offertes par l'édition électronique. De cette façon, on pourra espérer avoir des documents sources offrant un véritable potentiel d'exploitation du médium électronique.

L'expérience acquise au cours de ce projet a montré que notre modèle était bien adapté aux revues en transit vers l'électronique. De plus, de nombreuses applications SGML (en particulier les *Interactive Electronic Technical Manuals*[19]) ont montré que ce modèle pouvait très bien fonctionner pour des publications purement électroniques et très sophistiquées. Par conséquent, l'investissement dans un tel modèle risque d'être payant, car il pourra accompagner la revue tout au long de son existence, en s'adaptant aux nouvelles technologies ainsi qu'aux différents intervenants humains.

Sources consultées

Boismenu, Gérard, Martin Sévigny, Marie-Hélène Vézina et Guylaine Beaudry. 1999. *Le projet Érudit: Un laboratoire québécois pour la publication et la diffusion électroniques des revues universitaires.* Rapport sur le projet pilote réalisé par les Presses de l'Université de Montréal. Presses de l'Université de Montréal, juin 1999, 276 p. <URL:http://www.erudit.org/erudit/rapport.html>

Bullock, Alison. 1999. La conservation de l'information numérique: ses divers aspects et la situation actuelle par Alison Bullock. *Flash Réseau* (Bibliothèque Nationale du Canada) n° 60. <URL:http://www.nlc-bnc.ca/pubs/netnotes/fnotes60.htm>

Dugand-Saenz, Martha et Philippe Verdret. 1998. Créer des IETM avec la technologie Web ou comment rendre votre HTML intelligent? *Document numérique* 2 (2): 131-144.

Harnad, Steve. 1990. Scholarly skywriting and the prepublication continuum of scientific inquiry par Stevan Harnad, <URL:http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad90.skywriting.html>

Kasdorf, Bill. 1998. SGML and PDF - Why we need both. *Journal of Electronic Publishing* 3 (4). <URL:http://www.press.umich.edu/jep/03-04/kasdorf.html>

Lieb, Thom. 1999. HTML, PDF and TXT: The format wars. *Journal of Electronic Publishing* 5 (1). <URL:http://www.press.umich.edu/jep/05-01/lieb0501.html>

Marcoux, Yves. 1994. Les formats normalisés de documents électroniques. *ICO Québec* 6 (1-2): 56-65. <URL:http://tornade.ere.umontreal.ca/~marcoux/grds/ico94.htm>

Odlyzko, Andrew. 1999. Competition and cooperation: libraries and publishers in the transition to electronic scholarly journals. *Journal of Electronic Publishing* 4 (4). <URL:http://www.press.umich.edu/jep/04-04/odlyzko0404.html>

19. Il s'agit de normes pour la production de manuels techniques très sophistiqués. Voir <URL:http://www.ietm.net/>.