

Enquête sur XML

vers une édition multisupport

Nicole Rodriguez

ADJOINTE AU CHEF DE SERVICE
CNDP, SPIN

Avec l'apparition du numérique, l'édition a subi un bouleversement important. Les acteurs du domaine ont d'abord diversifié les supports d'édition pour y intégrer les supports électroniques, puis ils ont progressivement opté pour le format XML, qui assure la pérennité des contenus, quel que soit le support sur lequel le produit est ou sera diffusé. Histoire d'une mutation, en deux volets.

1. Les évolutions de l'édition depuis 1980

Les années 80 marquent l'entrée dans une nouvelle période, qui se caractérise par deux événements : la diversification des supports d'édition et l'arrivée de technologies numériques dans les chaînes éditoriales.



Les technologies numériques apparues depuis une vingtaine d'années ont fait émerger des évolutions importantes dans les activités éditoriales : édition sur Internet, ouverture au grand public de bases documentaires jusque-là réservées à des spécialistes de la documentation, productions multimédias interactives... Les filières d'édition sont cependant restées cloisonnées, les métiers étanches, étroitement accrochés au support de diffusion cible du produit final.

C'est ce que montre l'exemple du CNDP, éditeur de longue date de documents pédagogiques sur supports imprimés et audiovisuels, qui a introduit, au cours des vingt dernières années, les technologies numériques dans ses chaînes de production, en généralisant l'utilisation de la PAO pour l'imprimé, et en ouvrant prudemment sa production vers des supports électroniques (disquettes d'abord, puis CD-Rom, Web, et maintenant DVD-vidéo).

C'est ainsi qu'on parle aujourd'hui d'édition imprimée, audiovisuelle, en ligne, multimédia hors ligne... On est sur une conception de l'édi-

XML, c'est quoi?

XML (*eXtensible Markup Language*) est un format standard ou plus exactement un métalangage de balisage. Dérivé de SGML, modernisé et simplifié pour le Web, il permet comme SGML, par la pose de balises préalablement définies et adaptées au fonds documentaire traité :

– d'identifier et de décrire aussi finement que nécessaire, c'est-à-dire élément par élément, la structure d'un corpus de documents, en principe indépendamment de leur présentation visuelle;

– d'introduire pour chaque élément une description documentaire, appelée de plus en plus communément « métadonnées », appliquée au document;

– d'attacher au document des données multimédias sous différents formats (photos, vidéos, sons, musique, animations, simulations, pages web...) par l'intermédiaire d'un système d'ancrage contextuel;

– de proposer une navigation dans le corpus de documents par la pose de liens hypertextuels.

Seule la forme de la structure, arborescente, est imposée. Elle est définie par une DTD ou schéma XML, adaptée au type de document traité. Les bons outils de saisie ou de modification de documents en XML, en vérifiant que la structure est respectée, permettent d'obtenir un fonds éditorial et documentaire structuré.

Des outils adaptés au support assureront ensuite la mise en forme finale, en principe de manière automatisée, en s'appuyant sur la structure du document pour déterminer la représentation adéquate.

« ... des systèmes de publication dans lesquels on s'efforce de rendre les contenus éditoriaux indépendants de leur forme de présentation afin de pouvoir les distribuer sur plusieurs supports. »

tion déterminée par le support de diffusion du produit, conception en bonne adéquation avec l'état des technologies des années 80-95. En effet, cette époque s'est caractérisée par la transformation des modes de représentation et de traitement des données qui, d'analogiques, sont progressivement devenus numériques. La matière première des chaînes de production éditoriales (textes, images, sons, animations, vidéos, séquences interactives...) a ainsi changé de nature, avec la généralisation de l'utilisation du traitement de texte et le développement des appareils numériques pour la capture de l'image et du son. Nous avons vu également les transformations conduisant au produit final s'y effectuer par l'intermédiaire d'applications informatiques, restant elles-mêmes fortement dédiées au support : la PAO pour l'imprimé, les applications multimédias interactives pour l'édition CD-Rom, le montage pour l'audiovisuel, etc.

Années 2000: le virage du haut débit

La conception d'une édition déterminée par le support de diffusion tend à devenir obsolète. Elle disparaît au profit de concepts tels que *publication plurimédia* ou *multisupport*, *marché du « cross-média »*. On entend par là des systèmes de publication dans lesquels on s'efforce de rendre les contenus éditoriaux indépendants de leur forme de présentation afin de pouvoir les distribuer sur plusieurs supports. Pourquoi?

La période récente se caractérise par le développement des réseaux numériques de transport de l'information, en croissance accélérée avec l'extension des réseaux à haut débit depuis 2001-2002.

Ainsi, des produits numériques circulent directement de l'éditeur au consommateur. L'éditeur ne maîtrise plus les conditions de réception du produit : celui-ci sera vu sur des écrans de taille et de définition totalement différentes (de l'ordinateur au *wap*), traité par des logiciels plus ou moins rustiques (messagerie, navigateurs internet...); il sera lu après impression sur une imprimante de bureau ou publié dans un journal, un livre...

Les produits devront également être facilement accessibles sur Internet. L'éditeur a donc une tâche supplémentaire, celle d'organiser au mieux l'accès à l'information qu'il place sur ses sites. Il guide l'utilisateur à travers des listes arborescentes de choix commentés, à travers des scénarios de consultation de bases de données... Il conforte les techniques des moteurs de recherche plein texte par l'insertion de métadonnées dans ses documents. Il s'intéresse au

« XML, c'est quoi? », au format XML

Les balises (en rouge) définissent la structure du texte, qui s'offrira à diverses mises en forme à partir de ce code.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
```

```
<!DOCTYPE Encadre SYSTEM  
"C:\commun\RepXML\Encadre.dtd">
```

```
<Encadre Nom="XML">
```

```
<NoteRedacteur>Ce qui suit est un encadré qui peut être  
isolé du reste du document, placé en hyperlien pour le  
Web, selon la charte graphique de la revue pour l'im-  
pression</NoteRedacteur>
```

```
<SujetEncadre>définition de XML(tentative)
```

```
</SujetEncadre>
```

```
<AuteurEncadre>Nicole Rodriguez</AuteurEncadre>
```

```
<TitreEncadre>XML, c'est quoi ?</TitreEncadre>
```

```
<Paragraphe>XML (eXtensible Markup Language) est un  
format standard ou plus exactement un métalangage de  
balisage. Dérivé de SGML, modernisé et simplifié pour  
le Web, il permet comme SGML, par la pose de balises  
préalablement définies et adaptées au fonds documen-  
taire traité :</Paragraphe>
```

```
[...]
```

```
<Paragraphe>Des outils adaptés au support assureront  
ensuite la mise en forme finale, en principe de manière  
automatisée, en s'appuyant sur la structure du document  
pour déterminer la représentation adéquate.</Paragraphe>
```

```
</Encadre>
```

concept de Web sémantique... Les métiers documentaires s'immiscent dans les chaînes éditoriales.

La forme et le contenu

Nombre de ces traitements sont rendus possibles par les technologies XML. Ainsi, la forme première de toute édition devient numérique... mais la forme première seulement¹!

En effet, si le « numérique » est un mode de codage de l'information, c'est également un ensemble de systèmes de traitement informatique et de dispositifs de transport de cette information. Ainsi retrouverons-nous nos livres, nos films, nos revues, nos pages web, nos morceaux de musique... en bout de chaîne dans le produit final qui, lui, existera sous des formes visuelles, audibles, interactives, différentes, sur un support traditionnel ou électronique.

La conséquence inéluctable de cette séparation de la forme et du contenu, c'est la transformation radicale des métiers de l'édition, qui devront distinguer très nettement, dans leurs processus de travail, la gestion des contenus intellectuels et éditoriaux des formes finales de représentation d'un produit, et intégrer complètement les métiers et savoir-faire documentaires dans leur chaîne éditoriale. La réussite de cette transformation dépend essentiellement des capacités de l'éditeur à en saisir l'intérêt et à suivre, de façon éclairée, raisonnable et maîtrisée, les évolutions des technologies de l'information et de la communication.

Un seuil déjà franchi par les éditeurs

Les éditeurs de presse sont les premiers à s'être engagés dans cette voie, saisissant l'opportunité économique de diversifier les modes de distribution des contenus qu'ils éditent : l'objectif principal d'un journal reste bien sûr la distribution dans les kiosques. Mais l'utilisation de la messagerie électronique et de sites internet permet d'atteindre un public désireux de disposer des articles dès la parution de la première édition du journal et quel que soit le lieu où il se trouve. La mise à disposition d'archives électroniques sur le Web, de services documentaires assurant des livraisons de sélections thématiques sur CD-Rom ou sur Internet diversifie la gamme d'utilisation du fonds éditorial. Ce faisant, la presse, secondée par d'autres éditeurs pionniers (d'encyclopédies notamment), a ouvert la voie et a permis le développement, maintenant exponentiel, d'outils (logiciels, normes et standards) assurant l'interopérabilité entre les systèmes de traitement et d'accès aux documents, la conservation des don-

Une chaîne Word-XML vers HTML ou la PAO

Fort de son expérience d'éditeur, le CNDP a développé une chaîne complète d'édition XML qui permet de générer, à partir de documents auteurs saisis sous traitement de texte (Word...), aussi bien des documents en HTML que des versions imprimables pouvant être reprises en PAO sous Frame Maker.

Se libérer de soucis techniques, libérer du temps

Chronologiquement, cette chaîne a été développée afin de pouvoir mettre en ligne rapidement de nombreux dossiers (que l'on trouve sur le site du CNDP) en libérant les équipes éditoriales et les intégrateurs HTML des soucis techniques et des temps de réalisation de la mise en forme finale. Il s'agissait de traiter toutes les semaines le magazine *Télédoc*, la collection nationale de dossiers en ligne *Thém@doc*, les *Mags* et autres publications.

Mise en page et mises au point

Les choix effectués pour la mise en place de cette chaîne de production Word-XML ont reposé sur l'insatisfaction engendrée par l'hétérogénéité tant des documents que nous mettions en ligne que des modes d'accès à ces documents. Nous sommes partis du fait que les contributions de nos auteurs nous parviennent dans la grande majorité des cas au format Word. Notre travail consistait alors en une mise au point éditoriale effectuée par l'équipe de rédaction (rédacteur en chef, chef de projet, secrétariat de rédaction). Il aboutissait à une mise en page finale du document adaptée au support de diffusion. Dans le cas de l'édition imprimée, ces deux opérations étaient conduites simultanément, et il était très rare de disposer d'une version validée du contenu indépendante de la mise en page finale, ce qui nuisait considérablement à la publication ultérieure sur d'autres supports, notamment sur le Web.

Dans le cas où la publication était destinée à l'imprimé, nous n'avions d'autre solution que de transformer le fichier issu de la PAO en fichier PDF compressé pour le mettre en ligne, avec tous les inconvénients que cela présente (poids du fichier, inconfort de la lecture sur écran...).

Versions déshabillées

Nous souhaitons également répondre aux besoins des utilisateurs en posant des métadonnées dans ces documents pour faciliter l'accès aux données par moteur de recherche (Spinoo ou autres...) et en proposant des versions imprimables de chaque dossier, déshabillées de la navigation spécifique à Internet (sommaire, liens vers des pages annexes...).

Après une période probatoire, il s'avère que ces outils fonctionnent bien. Les résultats obtenus dépassent les objectifs initiaux puisqu'ils ouvrent sur la PAO avec des modifications minimisées. Ces outils progressent évidemment pour répondre aux besoins nouveaux et aux évolutions technologiques.

Perspectives

Les études et outils réalisés par le CNDP arrivent dans une phase où la mutualisation et le partage commencent à être possibles, non seulement au sein du Scérén, mais aussi avec les partenaires publics et privés.

ENT et chaîne XML

La chaîne Word-XML -> HTML ou PAO est utile pour la création de documents par les enseignants et plus généralement les auteurs et acteurs des ENT, CRDP ou CDDP.

ENT et systèmes d'information documentaire, métadonnées

De même, les outils de pose de métadonnées, selon les recommandations basées sur une *Dublin Core* étendue aux nomenclatures documentaires préconisées par le CNDP, sont susceptibles d'être repris et généralisés dans une expérimentation ENT.

Spinoo et le Web éducatif

L'adoption de ces recommandations pour une période intermédiaire, dans l'attente de l'aboutissement des projets de norme AFNOR, serait souhaitable. En effet, des métadonnées structurées, même intermédiaires, peuvent être récupérées par des automates pour être converties en schémas stabilisés. Ces états intermédiaires amélioreraient considérablement l'appropriation actuellement hasardeuse de documents sur le Web éducatif, en attendant le nirvana futur ! La prochaine version de Spinoo, en préparation, prendra en compte l'existence de ces métadonnées.

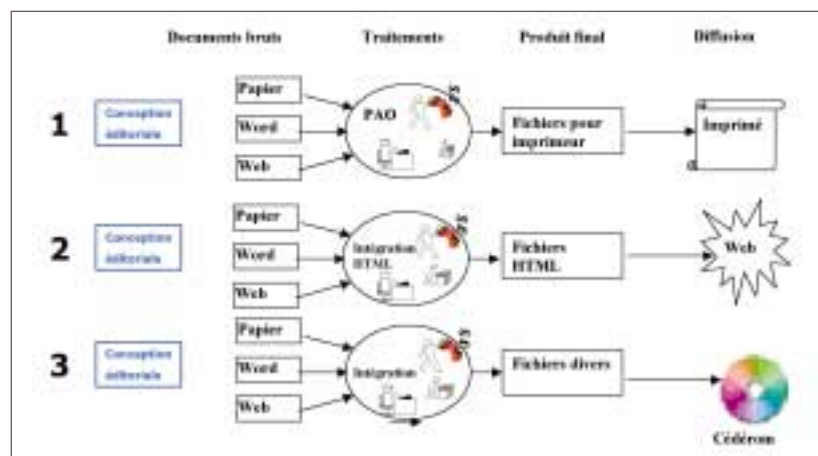
nées sous des formats pérennes, modifiables et interchangeables. ●

1. Réflexion inspirée par la lecture du document « Systèmes d'information et travail de groupe » présentant le projet Pelléas, Odile Arthur, CNRS, Christine Fabre-Broweys, UMLV. <http://cri.univ-mlv.fr/activites/bibli/intro-pelleas.pdf>

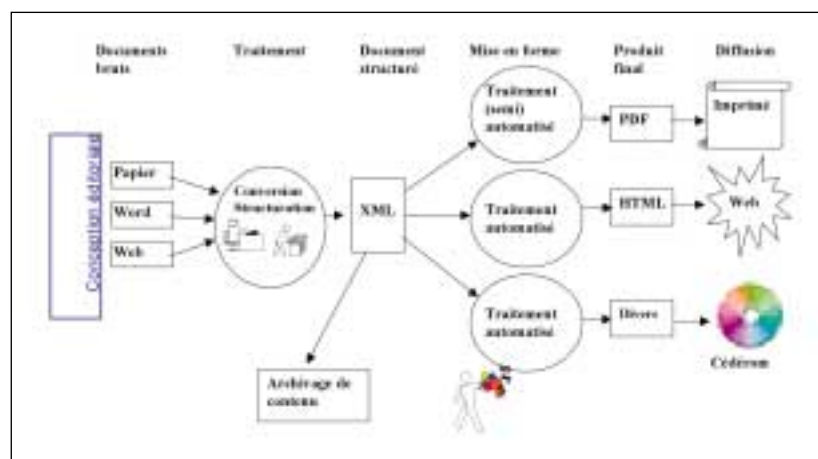
2. Redéfinition de la chaîne éditoriale

Méthodes de travail, constitution des équipes, métiers... les modifications introduites dans la chaîne éditoriale par l'émergence des technologies XML présentent certes des difficultés, mais marquent surtout des avancées vers une édition réellement multisupport.

Les différentes étapes vers la réalisation d'un produit ont des interactions étroites entre elles et ne se déroulent pas forcément linéairement. Dans les schémas ci-dessous, nous avons néanmoins conservé le découpage classique des chaînes éditoriales traditionnelles, en ne signalant par ailleurs que les faits nouveaux liés à l'introduction de XML.



Chaînes de production traditionnelles : trois chaînes différentes selon le support final envisagé.



Chaîne de production basée sur XML : une seule chaîne remplace les trois précédentes, puisque le contenu éditorial est traité indépendamment de sa forme finale.

Dès la conception de sa politique, l'éditeur devra prévoir pour chacune de ses lignes de produits (revues, collections, services...) le sort qu'il réserve aux corpus de contenus qu'il réunit. Trois questions à poser : la ligne de produits sera-t-elle déclinée sur plusieurs supports ou un seul ? les contenus prévus présentent-ils une certaine pérennité qui mérite de prévoir une réutilisation ultérieure dans des produits dérivés ? ces contenus auront-ils leur place dans un service internet ? Dès lors que la réponse à l'une de ces trois questions est positive, l'éditeur devra, simultanément à la réalisation de chaque produit, organiser la gestion des fonds éditoriaux qu'il réunit au fil de la déclinaison de l'ensemble de sa production.

Impact sur le processus d'édition d'un produit ou d'un service

Classiquement, une fois établie la charte éditoriale d'une ligne de produits, le processus de déclinaison de chaque titre suit un déroulé connu.

Conception éditoriale et élaboration des contenus

La conception éditoriale du titre est définie par une description détaillée des contenus souhaités et de l'organisation prévue (plan pour un ouvrage ou numéro de revue, interface d'accès et interactivité pour un CD-Rom ou un service internet...). Une équipe de projet est mise en place. Organisée autour d'un chef de projet ou d'un rédacteur en chef, elle suit la réalisation du produit : réunion et mise au point des contenus, maquette, production, test et fabrication. Elle fait appel au moment opportun aux compétences nécessaires : spécialistes de contenus, auteurs, secrétariat de rédaction, graphistes, réalisateurs de médias (sons, images, photos, animations, films...), informaticiens, spécialistes de bases de données, maquettistes, etc.

Dans le cas qui nous préoccupe ici – déclinaison sur plusieurs supports, gestion de fonds éditoriaux pérennes en vue de leur réutilisation –, plusieurs autres éléments devront être introduits.

D'abord, il faudra analyser et établir la structure logique des contenus textuels traités indépendamment de toute présentation visuelle. Introduction, parties, chapitres, titres, sous-titres, chapô, intertitres, résumés, encadrés, notes, insertion d'illustrations et d'éléments médias, légendes, copyright, auteurs, qualité des auteurs, etc. seront identifiés

et décrits. Cette structure sera traduite en une DTD XML structurant le contenu éditorial du texte et valable pour une ligne de produits donnée.

Il faut ensuite introduire des métadonnées de descriptions documentaires des contenus traités : sujets, niveaux, disciplines, indexation par thésaurus, etc., afin d'assurer la gestion du fonds documentaire, et l'accès aux documents eux-mêmes.

Enfin, il faut prévoir l'insertion d'informations liées à la production finale induite par le support, telle, par exemple, la description des droits d'utilisation d'une image sur chacun des supports ciblés, ou encore les liens hypertextuels entre entités du corpus... Ces éléments peuvent être consignés selon les choix effectués en bases de données ou par DTD XML.

L'adjonction de compétences documentaires à l'équipe de projet est donc indispensable, on le comprendra aisément.

La gestion du fonds numérique, la préproduction

On bascule du côté de la réalisation informatique pour laquelle un chef de projet technique, des développements et l'utilisation astucieuse de produits du marché seront nécessaires.

Numérisation

Rien de nouveau, sinon qu'il faudra prendre soin d'adopter des formats standard et non propriétaires pour les données multimédias afin de pouvoir les exploiter sans difficulté sur plusieurs supports.

Structuration des données

Il faut organiser la saisie structurée en XML et/ou en base de données des contenus éditoriaux et documentaires, en utilisant soit des outils du marché, soit des outils développés sur mesure.

Les outils du marché adaptés à la saisie de documents structurés, en XML ou en bases de données, ont une ergonomie contraignante, loin des habitudes prises avec les outils de traitement de texte. Pour des saisies de volume important, il sera utile de faire appel à des compétences spécialisées dans l'emploi de ces outils.

Une autre solution est de mettre en place des chaînes de saisie plus « douces », constituées d'une suite de transformations dans la mesure du possible automatiques, effectuées à partir de saisies sous traitement de texte.

Accès aux données, maquettage et automates de prévisualisation

Des scénarios et des automates de prévisualisation des données aussi proches que possible des formes finales devront être réalisés afin de permettre aux équipes en charge du contenu d'y accéder. Graphistes et informaticiens s'associeront pour la réalisation de ces automates.

La mise au point et la validation des contenus

Il serait souhaitable d'effectuer avant la mise en

forme finale le travail de corrections, remaniements, coupures, ajouts... aboutissant au calage définitif du contenu. On se heurte néanmoins fortement aux traditions de l'imprimé, où la validation finale par les auteurs et responsables de contenu n'intervient que sur la mise en page définitive.

Arrivé à ce stade, l'éditeur dispose d'un fonds éditorial validé, structuré et documenté en XML, indépendant de la représentation visuelle exigée par le support de diffusion.

Production et déclinaison multisupport

La composition, la mise en page, la navigation et l'interactivité, adaptées en fonction des supports de diffusion, imprimé, écran d'ordinateur, messagerie, PDA..., seront assurées par des automates qui s'appuieront sur la structure pour réaliser la présentation du contenu. Notons que la composition automatique, en particulier pour les versions imprimées, n'est pas toujours simple, ni possible. Des interventions de graphistes-maquettistes pour la mise en page finale sont alors à prévoir, ce qui serait facile si l'ensemble des outils de PAO intégrant complètement et correctement les technologies XML. Ce n'est malheureusement encore pas le cas pour la plupart d'entre eux.

Et si TDC avait franchi le seuil XML

Imaginons que la revue *TDC* soit publiée depuis dix ans selon ces méthodes. Le CNDP aurait accumulé un important fonds éditorial et documentaire dans le flux de sa production imprimée bimensuelle. Il serait donc à même de l'exploiter en créant des produits dérivés, sur des supports différents. Par exemple, un service d'archives sur Internet, un titre sur CD-Rom dédié à un sujet donné suivi sur plusieurs années, ou encore un recueil imprimé destiné aux parents d'élèves du primaire. Les métadonnées décrivant les articles lui serviraient à établir des scénarios de navigation. Comme les textes seraient conservés indépendamment de leur représentation finale, ils seraient aisément modifiables : actualisation de certains articles périmés, ajout de commentaires qui éclaireraient des articles anciens en fonction d'événements récents, placement de liens hypertextuels entre articles qui se compléteraient, introduction d'éléments multimédias (cartes, images, simulations, extraits vidéos...) qui animeraient le fonds...

En attendant que le rêve devienne réalité, nous avons malgré tout rétroconverti en XML le fonds de *TDC* sur les dix dernières années, en prenant deux articles seulement par numéro et en nous privant des images illustratives (pour lesquelles les droits d'utilisation sur Internet auraient été exorbitants) afin d'ouvrir prochainement un accès à ces archives sur le site du CNDP. ●