EXPERTISE TECHNIQUE ET ORGANISATIONNELLE POUR LES REVUES NUMÉRIQUES

GUYLAINE BEAUDRY
GÉRARD BOISMENU

G. Beaudry et G. Boismenu,
« Expertise technique et organisationnelle pour les revues numériques »,
dans site web *Expertise de ressources pour l'édition* de revues numériques,
sous la direction de G. Chartron et J.-M. Salaün, avril 2001,
URL : http://revues.enssib.fr/Index/indextecnic.htm.

EXPERTISE TECHNIQUE ET ORGANISATIONNELLE POUR LES REVUES NUMÉRIQUES

DE LA REVUE PAPIER À LA REVUE ÉLECTRONIQUE

Le défi de la publication électronique peut être vu comme une importante occasion à exploiter pour la revue. Certes, on doit y voir un moyen pour améliorer ses conditions de diffusion, mais c'est aussi progressivement une formidable situation qui incite à repenser la « forme » revue, ainsi que le mode de formalisation et de communication des résultats de recherche.

Nouvel environnement et revue savante

Le défi consiste à s'approprier les technologies de l'information afin de les mettre au service de la communication scientifique. La transition touche tout autant la valorisation et l'exploitation des résultats de recherche, la communication des travaux, les instruments de recherche et le rôle de l'éditeur que les composantes de la chaîne de la communication scientifique — partant des auteurs, en passant par l'éditeur et les bibliothèques, jusqu'au chercheur-lecteur.

La diffusion électronique des revues savantes et des autres genres de documents universitaires font en sorte que le repérage et la consultation de la documentation de recherche procèdent, pour une bonne part, selon des conditions entièrement renouvelées. Ce nouvel environnement a un impact majeur sur les pratiques d'exploitation et de formalisation des résultats de recherche dans la conception et

la rédaction des textes (articles, thèses, communications, etc.). Les auteurs-chercheurs doivent se doter d'habiletés indispensables permettant de connaître et même d'exploiter les fonctionnalités offertes par le traitement électronique des textes, afin de participer au bouleversement de l'univers de référence dans le domaine de l'exploitation et de la diffusion des résultats de la recherche.

La transition à l'électronique sera sur la bonne voie lorsque, avec ou sans version imprimée conventionnelle de la revue en parallèle, la version électronique du texte sera considérée première, et conçue comme telle, et non plus une simple vitrine électronique d'un document imprimé, écrit en intériorisant les contraintes du papier. On ne peut pourtant se cacher qu'actuellement la version électronique de la revue est d'abord une translation de la version imprimée vers le numérique. C'est que, tout en étant dans un nouvel environnement, les pratiques de rédaction et d'édition restent assez conventionnelles. D'ailleurs, plusieurs revues uniquement électroniques adoptent un comportement peu original eu égard aux usages confirmés dans l'environnement imprimé.

Une réflexion sur la façon de concevoir la revue, sa forme d'existence et les pratiques que cela peut induire devrait s'imposer. Le chercheur et la revue savante sont appelés à participer au renouvellement des conditions d'élaboration et de transmission des contenus par lesquels ils communiquent les résultats de recherche. Cette réflexion, et les innovations qu'elle peut susciter, dépendent de la sensibilité des directions de revue, mais certainement des secteurs disciplinaires et des types de document que cela suppose. C'est ainsi que, selon les secteurs, on jugera avec plus ou moins d'intérêt la possibilité d'intégrer dans le corps de la revue des éléments multimédias, des liens hypertextes, des données dynamiques, etc.

L'introduction de l'électronique dans les activités de publication devrait éventuellement participer à une réingénierie, tout autant des diverses étapes qui ponctuent le processus éditorial que de la façon de travailler, depuis la soumission des articles jusqu'à leur diffusion en version électronique. L'effet de l'introduction des processus élec-

troniques ne sera certainement pas instantané et suivra des rythmes, difficile à prévoir.

Une transition

La question centrale c'est de savoir comment s'opérera la transition. Un ensemble de conditions sociales, institutionnelles, culturelles et techniques doivent être prises en considération.

Par exemple, la simultanéité des deux supports, papier et électronique, est une contrainte avec laquelle on devra souvent composer. À la lumière des expériences locales et étrangères, la transition des revues imprimées, déjà bien établies, passe par le maintien dans la publication, la diffusion et la conservation des supports papier et électronique. Il appartient aux divers acteurs gravitant autour des revues de déterminer la durée de cette période, si tant est que nous devions assister à la disparition du papier.

De là se pose la question de la finalité qui motive, aux plans institutionnel et intellectuel, l'adoption d'une version électronique de la revue. On peut esquisser deux cas de figure.

D'un côté, l'accroissement de la diffusion est la motivation principale. Cette motivation, qui n'est pas absente dans l'autre cas de figure, a nettement préséance sur d'autres dimensions. L'accent mis sur la diffusion peut être inspiré, par exemple, par la vision concernant l'intérêt de l'électronique dans la communication scientifique, ou par la rareté de ressources disponibles en expertise ou en équipement, ou encore par la volonté d'étaler dans le temps la transition, en ne modifiant pas, au moins provisoirement, la façon de produire la revue.

D'un autre côté, l'électronique est perçue à travers les possibilités offertes, tout autant pour la production que pour la diffusion de la revue et l'action est menée sur ces deux volets simultanément. L'introduction de l'électronique, dans le processus même de la publication de la revue, contribue à faire en sorte que la version électronique soit conçue comme la version première, tout autant par les auteurs, que par les éditeurs et par les lecteurs.

Entre ces deux cas de figure, on peut imaginer plusieurs scénarios et différentes façons de faire. Aux fins de ce document, nous mettrons l'accent sur la chaîne de traitement XML-intégrée, qui assure une transition maîtrisée à la fois des processus de publication et de diffusion, et qui offre les meilleures garanties pour la préservation des documents. Cette chaîne réunit, de plus, les conditions pour favoriser le passage à la version électronique comme version première de la revue. Que la chaîne XML-intégrée puisse être perçue comme un scénario optimal ne doit pas conduire à sous-estimer d'autres façons de faire ; c'est pour cette raison que nous allons considérer trois autres avenues, allant de la seule diffusion d'une version électronique — qui est une duplication de la version papier —, jusqu'à une version en langage structurée, qui conserve, en parallèle, une chaîne de traitement pour le papier.

L'introduction de la publication électronique a nécessairement des implications pour les revues, notamment, en ce qui concerne le travail éditorial et de préparation de la copie. Dans tous les cas, la responsabilité éditoriale reste tout aussi entière : les équipes de revues poursuivent les mêmes activités éditoriales et ont la charge de livrer aux responsables de la publication une copie, dont le contenu est prêt à publier. L'impact le plus significatif pour l'équipe se manifesterait par l'adaptation des procédures de travail et par l'acquisition d'une formation assez légère. Par ailleurs, il faut pouvoir s'assurer que les revues possèdent des équipements informatiques de bon niveau pour exécuter les tâches attendues. Ces deux volets, formation et acquisition d'équipement, ont une incidence financière dont il faut tenir compte. On retrouve cette préoccupation dans plusieurs projets de publication électronique ; on peut citer, à titre d'exemple, J-Stage : Electronic Journal Publication & Dissemination Center. (http://www.jstage.jst.go.jp).

UNE INFRASTRUCTURE

La présentation qui suit ne suppose pas de cadre organisationnel particulier. Elle s'adresse tout autant à une direction de revue qui entend s'engager dans l'édition électronique qu'à un promoteur de site collectif de revues. On ne peut ignorer que, dans un projet de publication et de

diffusion électroniques de revue, la nécessité de combler des besoins en ressources assez diverses est particulièrement exigeante pour une revue seule, bien que l'on ne puisse pour autant conclure que cela soit hors de portée. Sans préjuger de la nature de l'organisation, il importe de considérer les composantes de l'infrastructure à mettre en place et des fonctions qui doivent être remplies.

Des services

Plusieurs services sont supposés par un projet d'édition électronique de revue savante. L'énumération des principaux services sera suivie d'une mise en relief des composantes des infrastructures devant être réunies, tout autant au plan intellectuel, organisationnel que matériel.

Le service de publication reçoit les textes, prépare la copie, prépare les épreuves, numérise les figures ou photographies, fait la saisie des corrections, prépare la version électronique et la mise en page pour l'imprimé. Dans la foulée, les métadonnées sont générées, idéalement automatiquement, à partir de la chaîne de traitement.

L'intérêt d'une collection d'articles de revues savantes réside dans la qualité des articles, mais aussi dans le nombre de revues et d'articles sur un sujet donné. Une étude (Boismenu G. et Beaudry G., 1999) récente a permis de noter l'intérêt manifesté pour les articles parus jusqu'à dix ans avant le moment de la cueillette d'informations. Ce qui justifie la numérisation rétrospective des numéros antérieurs pour en faciliter l'accès et offrir une collection d'articles plus riche. Il s'agit alors de réaliser la numérisation des articles visés et de créer les métadonnées pour permettre aux lecteurs d'utiliser un même outil de recherche à travers les collections courante et rétrospective d'une revue.

La seule présence dans Internet d'une revue n'est pas synonyme d'une réelle diffusion électronique. Le système de diffusion consiste en une interface d'accès et de consultation pour les lecteurs ainsi que des différents mécanismes à mettre en œuvre pour relier les revues portant sur les mêmes sujets et diffusées dans différents portails, à travers le monde. L'objectif est de rendre accessible une collection d'articles non pas par frontières géographiques, institutionnelles ou même disciplinaires, mais bien par sujets. La diffusion consiste aussi

à porter une attention particulière aux performances de ses différentes pages dans les outils de recherche les plus consultés et à conclure des ententes pour assurer la diffusion des articles par les différentes bases de données bibliographiques qui permettent les liens avec le plein texte.

Les outils proposés pour la mise en œuvre des services de diffusion sont la création d'un portail, l'utilisation d'un modèle de métadonnées, l'utilisation d'un outil pour l'identification et le référencement permanent de chacun des articles (par exemple DOI [Digital Object Identifier] ou PURL [Permanent URL]) et éventuellement l'utilisation d'un service tel que CrossRef ou SFX permettant de relier les références bibliographiques aux articles cités. Ce service doit également s'assurer de diffuser les tables des matières des numéros de revues accompagnées des résumés aux personnes intéressées ou abonnées à la publication.

La responsabilité de l'archivage de la documentation électronique revient en dernier ressort aux bibliothèques nationales. Comme le modèle et les pratiques ne sont pas encore établis, les revues doivent se doter d'un service d'archivage qui corresponde aux normes fixées par les archivistes. Quel que soit le partage des responsabilités en matière d'archivage, la revue doit assurer la conservation de la version électronique de la revue pour en garantir l'accès pendant tout son cycle d'utilisation.

L'accès aux revues sous forme électronique peut être gratuit ou payant. Dans ce dernier cas, il faut voir à établir un service de gestion des abonnements.

Des ressources intellectuelles

Pour répondre à ces diverses fonctions, des ressources intellectuelles de premier ordre sont nécessaires. Lorsqu'on se situe à l'échelle d'une seule revue, ces ressources peuvent être concentrées dans une seule personne particulièrement éveillée et dont l'engagement ne fait pas défaut. À l'échelle de plusieurs revues ou d'un site fédérateur, une spécialisation des tâches est davantage possible où, dans la division du travail, des qualifications diverses sont réunies.

Quel que soit le cas de figure retenu, il est précieux de compter sur une connaissance assez fine du milieu des revues et de réunir des

compétences diverses afin de concevoir, d'implanter et d'assurer la prestation de services, en tenant compte de l'évolution des normes et pratiques ainsi que des ressources financières disponibles.

Cela implique une direction intellectuelle arrimée aux conditions de réalisation, une gestion professionnelle en phase avec un programme de « livraison » de services, soit des compétences professionnelles et techniques complémentaires. Quatre dimensions peuvent être mises en relief.

L'accroissement de la visibilité, l'augmentation du nombre d'utilisateurs et l'impact du site sur la communauté scientifique internationale varient en fonction de la capacité de créer un lieu intellectuellement dynamique, qui nourrit, accueille et provoque des activités en mesure d'interpeller le milieu des chercheurs nationaux et internationaux. La meilleure promotion pour le portail, c'est d'en faire un foyer intellectuel dynamique qui fait place à des activités et offre des services de premier ordre. Cela demande des ressources conséquentes en qualité et en importance.

La publication électronique est un domaine dominé par l'innovation : tant les technologies utilisées que les usages et les pratiques dans le Web évoluent très rapidement et interpellent régulièrement les attentes et les prévisions concernant la publication et la diffusion électroniques. On est loin d'un univers où les conventions relatives aux processus de publication et à l'interaction avec les utilisateurs sont stables. Les activités consacrées à la veille sont fondamentales compte tenu de l'évolution des pratiques et des normes en matière de formats, d'outils de production et de diffusion, et de définition et transformation des modèles de diffusion tant technique qu'économique.

La nécessité de se situer à la frontière technologique a pour contrepartie de considérer les ressources dans la recherche et le développement d'applications comme le « nerf de la guerre ». On ne peut échapper à la nécessité d'investir dans la recherche appliquée, dans l'expérimentation et dans l'amélioration des fonctions de diffusion, surtout si on veut participer à la stabilisation prochaine du processus de production et au développement de services qui révèlent tout l'intérêt de la publication électronique en en exploitant les fonctionnalités.

Dans ce contexte, il faut être conscient de ce qu'est la recherche : l'expression « petit r grand D » correspond davantage à cette mission, car l'essentiel porte sur le développement et l'intégration d'applications pour la production et pour la diffusion.

La grande mobilité des ressources humaines dans ce secteur rend nécessaire l'identification des conditions permettant d'assurer une redondance des expertises pour la gestion et le fonctionnement quotidiens des systèmes d'information. La documentation et la diffusion des pratiques et des décisions deviennent, dans ce contexte, très importantes. C'est en fait ce qui permet à tout organisme de mieux gérer les connaissances et les savoirs des forces vives de son personnel.

Un environnement technique et matériel

L'édition électronique des revues n'est possible que si l'on réunit les conditions pour assurer la sécurité, la stabilité et la constance des services, aussi bien pour l'hébergement des documents produits que pour leur diffusion et leur archivage. Ces caractéristiques sont à la fois techniques et organisationnelles.

Le site qui héberge une ou des revues contient des données sensibles et stratégiques ; il est recommandé qu'il ne soit pas installé sur un serveur partagé, où d'autres sites Web et d'autres applications pourraient résider et compromettre son fonctionnement. Le site devrait être installé sur sa propre machine. Les besoins et l'environnement du lieu d'hébergement, avec ses composantes matérielles et logicielles, se définissent en fonction d'une série de facteurs dont on peut faire l'énumération :

- estimation de l'espace disque nécessaire pour les contenus : les articles de revues savantes (publications courantes et numérisation rétrospective) et les textes en prépublication ;
- autres besoins d'espace disque : espace requis pour le système d'exploitation, les divers ensembles logiciels pour la diffusion ;
- capacité d'ajouter, au moins progressivement, diverses fonctionnalités pour le système de diffusion ; par exemple : un outil de recherche plein texte et exploitant des métadonnées, la person-

nalisation des accès des visiteurs en fonction de leurs intérêts et de leurs droits d'accès, l'enregistrement des habitudes d'usage et l'adaptation de l'affichage à la machine visiteuse ;

- la technologie sous-jacente à l'architecture du site doit utiliser les technologies récentes les mieux éprouvées (Java/servlets, architecture ouverte, multi-plateformes, RDF, etc.).

L'hébergement devrait comprendre, en particulier, l'accès du serveur à une très large bande passante et une position stratégique sur les circuits d'interconnexion d'Internet, la surveillance continue des équipements et la continuité du service. Il devrait comprendre également la surveillance des appareils clients, la climatisation et l'alimentation garantie (UPS et génératrice), la prise de copie de secours, la maintenance des DNS pour établir les noms logiques de serveurs.

Cet environnement technique vient compléter les ressources humaines qui seraient mobilisées et il est à leur service. Il constitue une condition nécessaire de la mise en place d'une structure de production, de diffusion et d'archivage des revues électroniques.

LA PRODUCTION D'UNE VERSION ÉLECTRONIQUE

Un tour d'horizon des pratiques sur le plan international montre une variété dans les choix qui sont faits pour publier électroniquement une revue savante. La production de la version électronique d'une revue savante allant bien au-delà de la simple mise à disposition d'articles en ligne, pour chacune des trois fonctions — production, diffusion et gestion — une pluralité de pratiques est observée, que ce soit pour les formats de production ou les choix en ce qui a trait à l'archivage des documents électroniques ou aux modèles économiques.

Formats de production, de diffusion et d'archivage

Avant de passer à la production proprement dite, il importe de définir les outils avec lesquels nous allons travailler. L'édition électronique amène plusieurs distinctions relativement à l'édition papier. Une des distinctions fondamentales est que l'encodage numérique permet une multitude de transformations du contenu d'un document. Alors que le

document papier est fixé au terme de sa production et qu'il demeurera tel pour toutes les autres étapes de ses cycles d'utilisation, le document électronique se voit prendre plusieurs formes, que ce soit au moment de la production, de la diffusion ou de la gestion et de l'archivage. Le document électronique est multiforme. Tantôt il est utilisé pour la diffusion dans le Web, tantôt il sert à créer les films et les plaques pour l'impression des exemplaires papier ou encore son contenu est exploité pour créer des métadonnées ou une base de données. Il importe donc, d'une part, de bien distinguer formats d'acquisition, de production, de diffusion et d'archivage et, d'autre part, de déterminer quels sont les formats à utiliser pour chacune de ces étapes du cycle de vie d'un document comme l'article de revue savante.

Dans le cas qui nous occupe, les formats d'acquisition sont, pour les documents textuels, les fichiers de traitements de texte commerciaux, le plus souvent Word. Il s'agit du format utilisé par le secrétariat de la revue pour l'étape d'uniformisation des textes (références bibliographiques, corrections linguistiques, etc.). Du format d'acquisition, on passe à un format de production, XML ou QuarkXPress par exemple. Une fois toutes les corrections saisies et le contenu validé, la version finale servira également pour l'archivage. Pour cette étape, le format XML est nettement à privilégier comme il en sera question dans une autre section de ce texte. Les formats de diffusion électronique sont le plus souvent HTML et PDF. On peut penser que dans un avenir assez proche, le XML pourra être utilisé directement pour la diffusion.

En ce qui a trait aux figures et aux graphiques, les revues les reçoivent dans une foule de formats, allant du fichier électronique en format vectoriel jusqu'à une mauvaise reproduction par photocopieur d'une gravure. Dans la mesure du possible, les fichiers de documents iconographiques doivent être conservés dans un format reconnu pour l'archivage, le format TIFF par exemple. Les formats de diffusion pour le Web sont, pour le moment, le plus souvent GIF ou JPEG.

Les formats structurés (SGML et XML)

Les formats structurés, en plus d'encoder le contenu d'un texte, rendent lisible par l'ordinateur la structure sémantique et hiérarchique d'un docu-

ment. Chaque élément d'un texte encodé à l'aide d'un langage structuré se voit attribuer des balises qui le délimitent et l'identifient. Par exemple, un titre sera encodé comme ci : <titre> Ceci est un titre </titre>. Par ce balisage structuré, le contenu d'un document est nettement distinct des diverses représentations qu'on peut en faire. Cette particularité fait en sorte que l'apparence ou la mise en page d'un document n'est pas encodée à même le contenu, comme c'est le cas pour les formats de traitement de textes propriétaires. Une application et un fichier distincts déterminent les attributs de style, de mise en page ou d'affichage de l'information.

Les documents structurés permettent l'échange et la réutilisation de textes numériques tout en préservant le contenu, les données et la structure sémantique d'un document des différentes utilisations qu'on fera de l'information dans le présent et le futur. Chaque élément d'un document structuré peut être stocké, recherché, réutilisé, extrait pour créer un autre document ou une base de données. Les deux formats de balisage structuré les plus utilisés sont le SGML et le XML.

Le SGML (*Standard Generalized Markup Language*) est un langage structuré normalisé par l'ISO en 1986. Le SGML est un métalangage qui permet de décrire la structure logique d'un document. Le cœur d'un système SGML, la DTD (définition de type de document), est la « grammaire » d'un genre de texte (article, livre, dictionnaire, etc.) dans laquelle on retrouve la description des éléments, de leurs contenus et des relations entre les éléments. SGML est rapidement apparu très intéressant pour la diffusion des contenus à la fois sur supports papier et électronique. Seulement, les investissements financiers ainsi que l'expertise requise pour la mise en place et la gestion d'un système SGML ont freiné l'implantation d'une telle solution dans certains milieux.

Pourquoi ne pas se limiter au HTML ? À première vue, le HTML (*HyperText Markup Language*) peut sembler être une solution intéressante. Ce format est une application simple du SGML. HTML est facile à apprendre et sa diffusion dans le Web ou sur d'autres supports, tels que le cédérom, se fait assez aisément. Cependant, on rencontre rapidement les limites du HTML. Le jeu de balises limité du HTML ne permet tout simplement pas d'identifier et de représenter adéquate-

ment les nombreux éléments souvent complexes d'une revue savante. Le HTML est encore pour un certain temps le format de diffusion le plus utilisé dans le Web, mais, compte tenu de ses limites, il n'est définitivement pas adéquat pour l'édition de revues savantes.

Le XML : solution entre le SGML et le HTML ? Le XML (*eXtended Markup Language*) a depuis 1998 le statut de recommandation du W3C (*World Wide Web Consortium* ; http://www.w3c.org/). XML est un langage de balisage structuré, basé sur le SGML, développé pour pallier les limites du HTML sans pour autant posséder les difficultés d'application du SGML.

Le texte d'un article en XML est structuré de telle sorte qu'il peut être matérialisé par plusieurs médias avec un effort minimum : papier, Web, base de données, synthèse vocale, etc. Tout comme le format SGML, le XML permet de distinguer le texte et les données qu'il contient, des représentations visuelles qu'on lui donne, papier ou électronique, selon des besoins actuels et futurs.

Le XML, c'est aussi une famille de technologies. Le XLink (*XML Linking Language*), une proposition de recommandation du W3C depuis le 20 décembre 2000 (http://www.w3.org/TR/2000/pr-xlink-20001220/), est la norme qui décrit la façon d'intégrer des liens hypertextes à un fichier XML. Le XLink permet notamment de qualifier la nature des liens. Par exemple, on pourra, à partir d'une zone ou d'un mot sensible, accéder à la biographie d'un auteur, sa bibliographie, ses coordonnées ou son affiliation. Le XSL (eXtensible Stylesheet Language) est un langage qui permet de développer des feuilles de styles pour la représentation à l'écran des documents. Il est basé sur le XSLT (*eXtensible Stylesheet Language Transformation*), un langage de transformation qui permet de faire la conversion d'un document XML à un autre type de document XML. Dans un système XML, les schémas permettront aux utilisateurs de développer leurs propres applications XML. Finalement, le XHTML (*eXtensible Hypertext Markup Language*) est une reformulation du HTML 4.0 en XML, en quelque sorte une passerelle pour favoriser le passage du HTML au XML.

Le PDF et le Postscript

Les formats de description de page, bien que diffusés dans le Web, sont surtout destinés à l'impression. Les deux formats de loin les plus utilisés sont Postscript et PDF (Portable Document Format).

Postscript, introduit en 1985 par la compagnie Adobe, a été conçu pour faciliter l'impression de documents, peu importe les environnements logiciel et matériel. Un fichier Postscript fait la description des différents éléments de la page, graphiques ou typographiques. Il n'y a aucune organisation structurée de l'information dans un document Postscript et les ajouts de liens hypertextes sont impossibles. Postscript est souvent utilisé comme format de transmission pour les ateliers d'impression.

Le PDF (Portable Document Format) est un autre format propriétaire introduit par Adobe. Grâce au logiciel distribué gratuitement par Adobe, Acrobat Reader, il peut être lu sur plusieurs plates-formes (Acrobat Reader est disponible pour Windows, Mac OS, UNIX et LINUX.). Ce format permet l'intégration d'hyperliens, de signets et de métadonnées. Le PDF a pour avantage de préserver l'apparence originale du document ; il est, de plus, habituellement facile de produire un document PDF dans un environnement contrôlé. Toutefois, le PDF est davantage un format d'impression sur demande qu'un format de visualisation ou d'exploitation. Format attrayant pour sa simplicité, il est souvent facile de créer un fichier PDF à partir de n'importe quelle application permettant d'imprimer un fichier. Outre le fait que PDF soit un format propriétaire — ce qui ne donne aucune assurance quant à la pérennité de l'information —, ce format ne fait que restituer à l'écran ce que l'on retrouve sur papier. Devant un document PDF à l'écran, on se retrouve devant un papier de verre. De plus, comme l'information d'un document PDF n'est pas structurée, les possibilités de recherche sont moindres.

On peut dire en ce sens que le PDF facilite la transition du papier à l'électronique. Il s'agit ici de l'application d'une technologie nouvelle à des outils d'un autre âge. C'est peut-être le propre des périodes de transition de procéder ainsi comme le laisse entendre Febvre et Martin dans *L'apparition du livre* :

> *[...] les premiers incunables ont exactement le même aspect que les manuscrits. En cette période de début, les imprimeurs, bien loin d'innover, poussent à l'extrême le souci de l'imitation : la Bible de 42 lignes, par exemple, est imprimée dans des caractères reproduisant très fidèlement l'écriture des missels manuscrits de la région rhénane. Longtemps les typographes utilisent non seulement des alphabets de caractères isolés, mais aussi des groupes de lettres liées entre elles par les mêmes ligatures que dans l'écriture manuscrite. (Lucien Paul Victor Febvre et Henri-Jean Martin, 1971, p.111)*

Les formats images

Une discussion détaillée des contenus multimédias dépasse le cadre de ce document. Toutefois, soulignons quelques principes qu'il faut respecter.

La question des formats se pose tout aussi naturellement aux contenus multimédias qu'aux contenus textuels. Ces formats doivent être choisis en fonction de la pérennité de l'information et de la structuration. Les formats permettant d'inclure des métadonnées dans les documents multimédias sont à privilégier, afin de faciliter leur repérage et leur manipulation. Cette pratique est encore peu courante. Les différents standards en cours de développement sont à suivre.

Un document encodé en format image est une représentation assez près d'une photographie d'un texte, d'une image ou d'un objet. En ouvrant un tel type de fichier, nous avons bel et bien à l'écran l'image d'un texte, par exemple. Cela implique que pour tous les formats images, aucune manipulation de texte n'est possible, que ce soit « copier-coller » ou encore la recherche plein texte.

Il existe deux grandes familles de formats images : (1) les formats vectoriels et (2) les formats image en mode point (Creating Digital Ressources for the Visual Arts : Standards and Good Pratice, 2000).

Les formats vectoriels décrivent l'image en une série d'objets géométriques (lignes, ellipses, polygones, etc.) dont les propriétés permettent de recréer l'image à partir d'instructions précisées. L'information contenue dans le fichier d'un format vectoriel décrira, par exemple,

la position, l'épaisseur et la couleur d'une ligne à tracer. Il s'agit d'une série d'instructions qui est réutilisée par l'outil de visualisation pour reconstruire l'image. Ce type de format permet d'éditer les différents objets d'une image indépendamment : les images peuvent être modifiées sans perte de résolution et la taille des fichiers est relativement petite.

Les fichiers image en mode point sont faits d'une mosaïque d'éléments d'image appelés pixels. Chaque pixel contient les informations concernant la couleur d'un point de l'image en particulier. La combinaison de tous les pixels forme l'image. Les fichiers image en mode point sont habituellement très lourds à cause de la quantité d'information que contient chaque pixel et du grand nombre de pixels nécessaires pour obtenir une image de qualité. Pour ces raisons, un format vectoriel est souvent à privilégier par rapport à un format image en mode point, sauf lorsque le type d'information ne s'y prête pas.

Comme en témoigne le nombre de formats différents qui sont acheminés aux rédactions de revue, une multitude d'outils sont utilisés par les auteurs et leurs assistants pour la production des images. Pour ce qui est de la diffusion électronique, les formats les plus utilisés sont GIF et JPEG. Le GIF (Graphics Interchange Format), créé par CompuServe en 1987 et amélioré en 1989, est un format très utilisé dans le Web. La compression du format GIF limite à 256 couleurs ou tons de gris, ce qui offre comme avantage de générer des fichiers très légers. Le GIF est utilisé en général pour les figures.

Le JPEG (Joint Photographic Expert Group) permet l'utilisation de 16 millions de couleurs et est surtout employé pour représenter des photographies. C'est également un format très répandu dans le Web, particulièrement reconnu pour son excellent algorithme de compression.

Le format TIFF (Tagged Image File Format) a été développé par Aldus et Microsoft. Le principal avantage du TIFF est son algorithme de compression qui assure aucune perte d'information. Ce format de type image en mode point est reconnu par les archivistes et plusieurs guides des meilleures pratiques pour l'archivage (Creating Digital Ressources for the Visual Arts :Standards and Good Pratice, 2000). Les fichiers TIFF sont particulièrement lourds avant compression et

ne peuvent être lus par les navigateurs Web. Ce format est surtout utilisé comme format de capture dans les projets de numérisation rétrospective.

Le développement et l'utilisation de nouveaux formats d'images doivent être suivis attentivement au cours des prochaines années. Par exemple, le SPIFF (Still Pictures lnterchange File Format) a été commandé par l'ISO afin de garantir l'interopérabilité des systèmes manipulant les images (Lecomte D. et al., 1999, 91.). Ce format, créé par le Joint Photographic Expert Group, fait partie du domaine public. Il s'agit d'une amélioration de leur norme JPEG. Le PNG (Portable Network Graphics; « A Basic Introduction to PNG Features », 2000) est la nouvelle étoile filante des formats d'images. Il est destiné à remplacer le GIF et le TIFF. Il s'agit également d'un format non propriétaire.

Il est possible d'obtenir des fichiers en PDF Image lors de la numérisation d'une image (Cleveland G., 1999). Pour ce faire, il faut utiliser le logiciel Adobe Capture. Ce format peut être utile pour les projets de numérisation rétrospective de revues savantes, comme on le verra un peu plus loin.

Chaînes de traitement : différentes options

Nous avons constaté d'entrée de jeu que plusieurs façons de faire étaient à l'œuvre dans la publication des revues. Tout en reconnaissant que le scénario XML-intégrée est sans doute le plus porteur et systématique pour la transition vers l'adoption de l'édition électronique, les autres façons de faire méritent une mise en situation. Précisément, ces différentes façons de faire se caractérisent diversement selon les enjeux de la publication électronique.

Pour simplifier le traitement, nous avons ramené les options à quatre (PDF-texte, HTML, Quark-XML, XML-intégrée) et donné schématiquement les caractéristiques de chacune des options pour les principaux indicateurs, regroupés par grandes fonctions de l'édition électronique des revues. Quelque 13 indicateurs sont énumérés et sur lesquels des précisions doivent être apportées.

La fonction production peut être plus ou moins étendu selon l'introduction des fonctionnalités de l'électronique. C'est pourtant là une condition indispensable pour assurer une transition faisant, à terme, de la version électronique de la revue la version première de référence — non seulement pour la consultation, mais bien en amont au moment de la rédaction et de l'édition. Cela favorise l'achèvement de la transition dans la mesure où, aussi longtemps que la version électronique sera dérivée de la chaîne de traitement pour le papier, l'environnement électronique restera tributaire de celui de l'imprimé.

- Cela a évidemment des conséquences techniques, mais surtout pose la question centrale de la capacité organisationnelle de mettre en place une chaîne de traitement en langage structuré (XML-intégré), fondamentalement basée sur un traitement électronique au début du processus de production dont les différents formats de diffusion découleraient. Les autres avenues sous-entendent un traitement préalable pour le papier, puis un traitement électronique. Dans ce cas, les distinctions se font sur la base du type de traitement électronique ; nous allons du traitement PDF-texte vers un traitement pour le Web, avec le HTML, ou encore en langage structuré, mais issu d'un traitement préalable sur logiciel de mise en page (d'où la mention Quark-XML).
- Ces choix conditionnent pour une bonne part les formats de production et de préservation, de même que les logiciels avec lesquels le travail est effectué. De façon générale, il est plus approprié, moins contraignant et plus économique d'utiliser des formats ouverts et des logiciels libres ou gratuits (pour le milieu de l'éducation, tout au moins). De là, la distinction basée sur le caractère propriétaire et non propriétaire des logiciels et formats.
- De même, le mode de production des métadonnées — qui décrivent les attributs et le contenu des articles et qui servent au repérage, à la gestion, à la description, à l'accès et à la conservation de l'information — doit, au mieux, procéder par extraction automatique de l'information des articles eux-mêmes, sinon cela

impose une procédure manuelle qui est fastidieuse et exigeante. D'où, la prise en compte de cette dimension.

Pour la diffusion, une série d'interrogations permet de qualifier les quatre options retenues.

- Selon le format de production utilisé, comme format maître, différents formats pourront éventuellement être générés. Dans certains cas, comme avec le PDF, cette possibilité n'existe pas, alors qu'à l'autre bout du spectre, l'utilisation d'un format riche, comme le XML, permettra de générer plusieurs formats dans une période donnée.
- Les formats ont aussi une incidence sur les capacités de recherche et les conditions pour améliorer l'efficacité des résultats, en termes de diminution du bruit et du silence. Sur ce plan, la possibilité de procéder à la recherche structurée constitue un avantage sur la seule recherche plein texte. Notons au passage que le PDF-image ne permet ni la recherche plein texte, ni la recherche structurée.
- Les besoins pour la diffusion des documents évoluent et se diversifient dans le temps. Plus le format de production est riche, plus la gamme des possibles s'élargit, ce qui constitue un avantage. Nous avons utilisé une variable dichotomique pour la réutilisation des données, en termes de facile et difficile.

Tant et aussi longtemps qu'une procédure fiable d'archivage des documents ne sera pas établie publiquement, l'éditeur de revue conserve une responsabilité à cet égard. D'une façon plus modeste, l'éditeur de revue, dans le contexte de la diffusion électronique, doit s'assurer de la continuité et de la fiabilité du service de mise en ligne, ainsi que de la disponibilité des articles. Cela demande nécessairement la préservation des documents, en dépit et au-delà de l'obsolescence technologique, tant pour les logiciels que pour les supports. Il est, en ce sens, primordial de tenir compte d'un indice de pérennité lié aux différents formats.

Les ressources qui doivent être mobilisées varient selon la chaîne de traitement retenue. Les coûts de production sont d'abord et avant tout fonction de l'expertise nécessaire à la réalisation du projet, alors

que les coûts en équipement (machines et logiciels) sont relativement semblables : les machines sont comparables et les logiciels comptent relativement peu dans l'ensemble.

Les quatre voies de transitions ont été schématiquement caractérisées pour chaque indicateur, ce que livre le tableau suivant. Les principaux constats qu'il contient sont repris dans la section suivante et intégrés à la présentation des grandes données concernant chacune de ces quatre voies de transition.

	PDF texte	HTML	Mise en page vers XML	XML intégré
Production				
• processus (électronique version première [1] ou dérivé le papier [2])	2	2	2	1
• format (non-propriétaire [1] ou propriétaire [2])	2	1 et 2	1 et 2	1 et 2
• logiciel (non-propriétaire [1] ou propriétaire [2])	2	1 et 2	1 et 2	1 et 2
• métadonnées (production automatique [1] ou manuelle [2])	2	2	1	1
Diffusion				
• formats (génération automatique de plusieurs formats)	non	non	Oui	Oui
• recherche (structurée [1] plein texte [2])	2	2	1 et 2	1 et 2
• réutilisation des données (facile ou difficile)	Difficile	Difficile	Facile	Facile
Préservation				
• indice de pérennité	Faible	Moyen	Élevé	Élevé
Ressources				
• coûts production	Faible	Moyen	Élevé	Élevé
• coûts équipement	Faible	Faible	Moyen	Moyen
• expertise	Faible	Moyen	Élevée	Élevée

Chaînes de production : modèles et enjeux

Quatre modèles sont présentés : PDF (Texte), HTML, Mise en page vers XML et XML-intégré. Le choix pour l'un ou l'autre doit être fait en fonction des implications techniques et organisationnelles de chacun en tenant compte des ressources et de l'expertise à la disposition d'une

revue ou d'un regroupement de revues. Le modèle XML-intégré est celui qui sera le plus développé pour des raisons tenant à la fois de son intérêt et de ses enjeux.

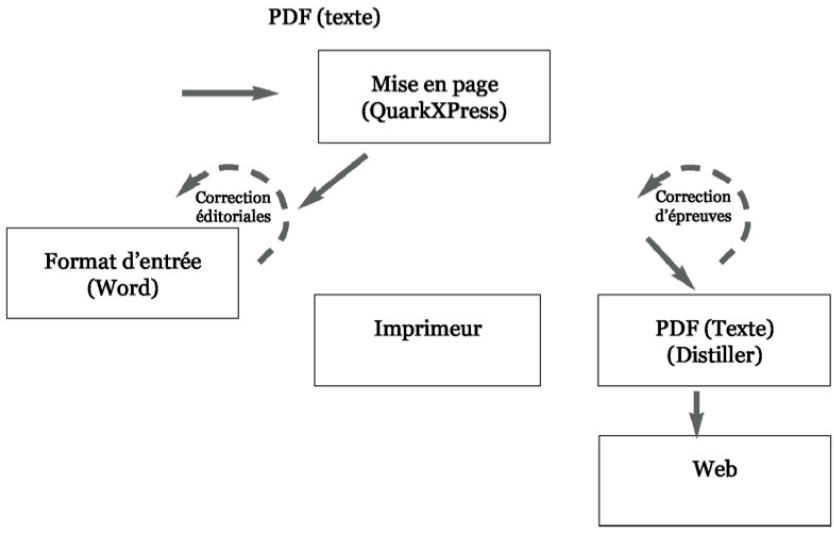
PDF (Texte)

Ce modèle est celui qui implique le moins de modifications à la majorité des processus de production. L'équipe de la revue reçoit des auteurs les articles en format traitement de textes, le plus souvent Word. Une fois les corrections éditoriales et l'uniformisation selon le protocole de rédaction terminées, la mise en page et les corrections d'épreuves sont réalisées à l'aide d'un logiciel de mise en page, tel QuarkXPress. De là, un fichier ou un prêt-à-photographier est produit pour l'impression des numéros.

Dans ce modèle, la production des fichiers PDF est tout simplement réalisée à partir de l'application logicielle de mise en page en utilisant le logiciel Distiller de Adobe. Cette opération se déroule le plus souvent sans problème, particulièrement lorsqu'elle s'effectue à partir de l'ordinateur où la mise en page a été réalisée, garantissant, par exemple, la présence de tous les fichiers de fontes utilisés.

Toutefois, ce modèle de production d'une version électronique, en s'appuyant sur le format PDF ne permettra que de mettre en ligne la translation de la version papier. Comme nous l'avons vu dans une section précédente, malgré les possibilités d'hypertexte du PDF, ce format est davantage considéré pour l'impression sur demande. De plus, compte tenu du fait que PDF est un format propriétaire, il est plus difficile de garantir les conditions de préservation à long terme. Des recherches en cours sur cette question identifient la conversion de PDF vers un autre format (image ou texte) comme la solution pour assurer la pérennité de l'information encodée en PDF (Ockerbloom, 2001). Finalement, PDF ne comportant aucune information structurée, la production des métadonnées ne peut se faire automatiquement et les possibilités de recherche sont limitées à la recherche plein texte.

Ce modèle est le moins coûteux à mettre en place, autant par l'équipement à acquérir, qui se limite à l'achat de Distiller, que par l'expertise nécessaire.



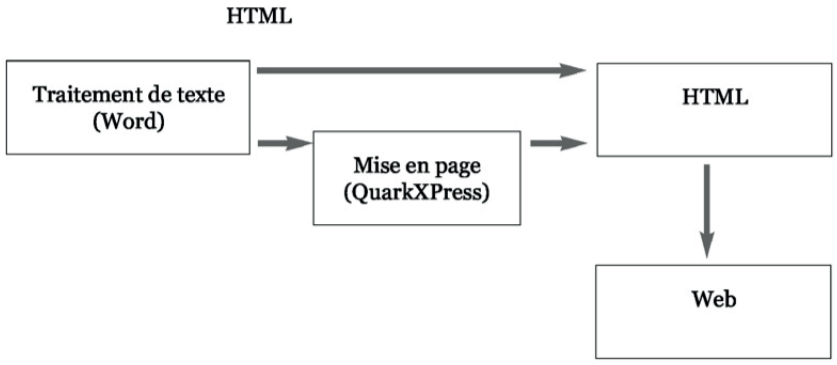
HTML

La plupart de applications de traitement de texte ou de mise en page permettent de produire une sortie en HTML. Cependant, l'utilisation de ces fonctionnalités automatiques donne souvent des résultats décevants. Le HTML ainsi produit est par définition calqué sur la représentation déterminée pour la forme papier. Il est alors nécessaire de retravailler les fichiers ainsi produits en élaguant les codes inutiles et en modifiant les balises automatiquement introduites. Cette procédure est évidemment fastidieuse et coûteuse. Il sera souvent plus opportun d'utiliser des outils qui permettront de mieux contrôler l'insertion des balises.

Le HTML, par son jeu de balises limité, ne permet pas de recherches structurées. Bien que des éléments comme « H1 » ou « Blockquote » puissent nous donner quelques indices, leur utilisation non-systématique et l'absence de balises pour identifier la majorité des éléments d'un article restreignent les possibilités de recherche et de représentation. Pour les mêmes raisons, la production automatique des métadonnées n'est pas possible. HTML étant une application du SGML et son jeu de caractères de base étant l'ASCII, une certaine pérennité peut lui être accordé. Toutefois, l'utilisation de balises propres à certains logiciels

et la juxtaposition des informations concernant la représentation de l'information (« Center » ou « Bold », par exemple) ne font pas du HTML un format à privilégier pour la préservation à long terme.

HTML est certainement un format de diffusion à utiliser pour encore quelques années. Cependant, les développements et l'utilisation de plus en plus courante du XML en font certainement un meilleur choix pour la revue savante.

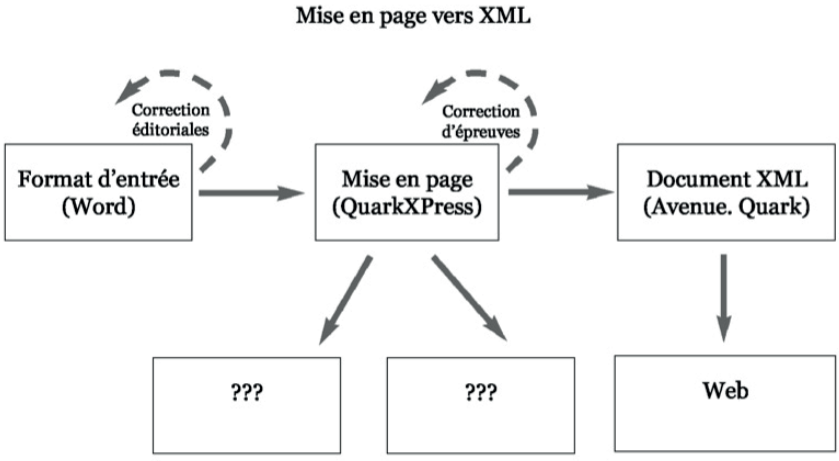


Mise en page vers XML

Les solutions PDF et HTML sont souvent jugées limitées pour répondre aux besoins de la publication, de la diffusion et de la préservation des revues sous formes numériques. Toutefois, l'implantation d'une chaîne de traitement XML intégrée permettant de bénéficier pleinement des possibilités du numérique, que ce soit pour la production de tous les formats de diffusion à partir d'un même document XML ou la réalisation d'une véritable revue électronique intégrant des éléments multimédias, implique des modifications importantes aux façons de faire. Pour des raisons tout à fait justifiées d'expertise requise ou d'organisation, l'implantation d'une chaîne de traitement XML-intégrée peut être difficile.

Dans cette situation, les possibles ne se limitent pas à des choix contraignants, c'est cet aspect que le présent modèle tente d'exploiter. La plupart des versions papier des revues sont produites à l'aide de

logiciels de mise en page, tels QuarkXPress ou PageMaker. Comme on le voyait précédemment, une sortie PDF à partir de ces applications est facile à produire. Toutefois, quelques-uns de ces logiciels permettent d'effectuer, à l'aide de module spécifique, des conversions de leurs formats natifs vers XML. Par exemple, le module Avenue. Quark du logiciel de mise en page Quark, à l'aide des styles appliqués au moment de la mise en page, assure par des règles la conversion vers XML.



Le premier désavantage de ce modèle tient à la dépendance pour un format propriétaire, tel que celui de Quark. Désavantage tout relatif, il faut bien l'avouer, puisque l'utilisation de logiciels propriétaires est courante et bien implantée pour la mise en page des revues savantes. Le deuxième inconvénient est que cette option fait du numérique un produit d'un processus conçu pour le papier. La chaîne de production est en fonction du papier, une fois la revue prête à imprimer réalisée, un fichier est à la fois acheminé à l'imprimeur et utilisé pour les opérations de conversion vers XML. Néanmoins, compte tenu de l'importance du papier, de l'expertise et de ces façons de faire éprouvées pour les équipes des revues, l'option « Mise en page vers XML » est certainement à considérer et plus facile à implanter qu'une chaîne de traitement XML-intégrée.

Pour les revues dont les auteurs écrivent en fonction de la diffusion papier, ce qui est le cas pour la majorité d'entre elles, ce modèle offre exactement les mêmes avantages que le modèle XML-intégré. Les variables touchant les possibilités de recherche, la génération automatique des métadonnées et l'indice de pérennité élevé sont du même ordre pour les deux modèles.

La chaîne de traitement XML-intégrée

Ce modèle est basé sur certains principes permettant d'atteindre les objectifs reliés à la publication de revues savantes dans un environnement électronique.

1. La normalisation. Cette chaîne de traitement utilise des formats normalisés pour la représentation de l'information et tente d'utiliser le plus souvent possible des mécanismes de traitement normalisés, sans mettre en péril l'efficacité et la souplesse de l'ensemble. La normalisation des formats d'encodage de l'information est nécessaire, car elle permet d'atteindre une certaine indépendance par rapport aux outils utilisés, mais aussi d'assurer la pérennité de l'information. Toutefois, aucune normalisation des outils n'est proposée, au contraire, l'indépendance par rapport aux logiciels et systèmes doit toujours être gardée.
2. L'exploitation de l'électronique. Les documents numériques provoquent des changements dans les façons de faire, tout en ouvrant des perspectives nouvelles pour la création, la diffusion et l'exploitation des résultats de la recherche. Ils permettent d'inclure différents types d'information, du texte à l'animation 3-D en passant par l'image, la vidéo, le son et la réalité virtuelle. Le mode de production d'une revue doit permettre l'exploitation du caractère multimédia et interactif de l'information électronique. Le processus de production d'une publication imprimée s'effectue, quant à lui, dans un contexte où l'on connaît et contrôle le support de diffusion de l'information. Dans le monde électronique, il existe une variété de formats de diffusion, mais surtout une grande variété de qualités ou de caractéristiques

techniques de ces formats. La chaîne de traitement décrite dans cette section répond au besoin de souplesse dans la production de l'information, électronique ou imprimée, afin de satisfaire le plus grand nombre de supports de diffusion, actuels et futurs, ainsi que le plus grand nombre d'expériences de consultation de l'information.

3. Une source, plusieurs produits. L'exploitation de l'électronique permet de fournir l'information aux lecteurs sous différentes formes, en différentes versions. La meilleure exploitation de la transition au numérique de la revue repose sur la modification du processus, ou sa ré-ingénierie, pour arriver à produire une source d'information unique qui sera déclinée en différentes versions.
4. Les documents structurés. Ceux-ci contiennent de l'information sur leur structure logique et sémantique. Par exemple, on y identifiera explicitement les différentes sections du document et leur titre ou encore certaines parties du texte comme des lieux géographiques ou des noms de personne. Cette structure leur confère une grande richesse, car il est aisé de traduire une information de structure en une information de restitution. Ainsi, convertir une information telle que « ceci est un titre de section premier niveau » en une instruction telle que « ceci doit être imprimé en caractères gras et un corps de 14 pts » peut se faire aisément et, surtout, automatiquement.
5. L'utilisation de documents structurés est la seule façon de respecter le principe « une source, plusieurs produits » dans un contexte électronique. Cette chaîne de traitement permet d'obtenir, pour chaque article, un document structuré, suffisamment riche pour permettre toutes les exploitations envisagées. Il existe deux normes permettant de représenter des documents structurés : le SGML et le XML.
6. L'automatisation. L'intérêt économique de la chaîne de traitement repose sur l'automatisation des différents traitements effectués sur l'information. La création de l'information demeurera toujours un processus intellectuel. Mais une fois cette information créée,

sa déclinaison en différents produits de gestion et de diffusion devrait toujours être réalisée automatiquement.

La production de l'information

La première étape de la chaîne de traitement consiste à obtenir un document électronique structuré, suffisamment riche pour servir de source unique aux multiples formes de publication et de diffusion. Le grand principe cher à tout éditeur de travailler le plus possible sur le manuscrit s'applique encore ici. Il devient essentiel que les équipes éditoriales des revues acheminent des textes dont le contenu est prêt à publier, de sorte que les corrections soient saisies une seule fois dans la source unique, pour ensuite servir à la production des différents formats de diffusion.

La chaîne de traitement utilise une approche indirecte et à deux étapes pour obtenir un document structuré XML qui constitue la source unique d'information. Dans une première étape, l'information est saisie (ou récupérée) dans un document de traitement de texte conventionnel (nous préconisons le traitement de texte Microsoft Word, le plus utilisé) en respectant des règles d'écriture assez strictes. Ensuite, ce document Word est converti automatiquement en format XML, sans aucun traitement manuel ou intellectuel.

Puisqu'il nous apparaît pour le moment improbable que la création des articles se fasse directement en XML, cette procédure permet de faire un compromis optimal entre la bonne connaissance du traitement de texte des différents contributeurs et la souplesse ainsi que la normalisation du format XML. Cette partie de la chaîne de traitement repose sur le respect des règles d'écriture du document Word qui constitue l'intrant de la chaîne de traitement. Pour profiter des avantages de la conversion automatique du format Word vers le XML, les corrections devront être saisies dans l'intrant de la chaîne de traitement pour produire, à partir du document Word, un contenu unique pour les versions papier et électroniques.

Le document Word

Le format Word (comme tous les formats de traitement de texte) ne se prête pas à la représentation de documents structurés, et ce, pour deux raisons principales :

- le format n'est pas hiérarchique, alors que les documents textuels le sont presque toujours ;
- les traitements de texte ne permettent pas de valider le contenu ou la structure des documents.

Ce dernier point est majeur, car une bonne utilisation des documents structurés implique la notion de validation (de la structure) assurant d'obtenir des documents à structure homogène et ainsi, l'exploitation de façon structurée. Il s'agit exactement du même principe que l'on applique aux bases de données, où l'on sépare en champs les différentes unités d'information, pour ensuite pouvoir les exploiter très facilement et efficacement. Les documents structurés appliquent ce principe aux documents textuels.

Même si le traitement de texte ne valide pas la structure d'un document, il permet de l'identifier en partie, et ce à l'aide des feuilles de styles. On peut facilement indiquer qu'un paragraphe est le titre d'une section de niveau deux en lui appliquant le style « Titre 2 », par exemple. Les styles de types caractères permettent d'obtenir le même effet sur quelques lettres ou quelques mots du texte, à l'intérieur d'un paragraphe.

L'application des styles doit nécessairement être effectuée par des personnes qui comprennent à quoi cette étape peut servir, qui connaissent les documents qu'elles s'apprêtent à « styler » et qui connaissent la feuille de styles à appliquer. Leurs connaissances de la revue et de leur discipline constituent la première validation.

Une version Word du document est produite en respectant un modèle établi spécifiquement pour la publication électronique sur plusieurs supports. La production de cette version « Word ++ » devrait être effectuée par les secrétariats des revues. La création de la version « Word ++ » est une tâche manuelle et intellectuelle qui consiste à baliser le document pour y ajouter de la valeur en identifiant la struc-

ture pour qu'elle soit manipulable par des outils informatiques. Une fois cette version « Word ++ » créée, une version XML est obtenue automatiquement. Le reste du processus consiste à créer des épreuves des articles, imprimées et électroniques, et à vérifier ces épreuves pour identifier et corriger les dernières erreurs. Les épreuves sont produites automatiquement, avec les différents outils de la chaîne de traitement qui seront présentés plus loin. La saisie des dernières corrections dans la version « Word ++ » terminée, le document XML de référence sera produit et, de là, toutes les versions de diffusion. Une validation humaine n'est toutefois pas totalement fiable, et c'est pourquoi une validation informatique est nécessaire. Celle-ci sera obtenue lors de la conversion du document Word au format XML, ce dont nous parlerons dans la prochaine section. Le point le plus important de cette partie de la chaîne consiste en l'équivalence des versions « Word ++ » et XML des articles, équivalence obtenue par la réalisation d'un puissant outil de conversion et qui permet de tirer parti des forces de chacun des formats, soit l'utilisation presque généralisée dans le cas de Word et la pérennité et la facilité de manipulation du format XML.

Le document XML

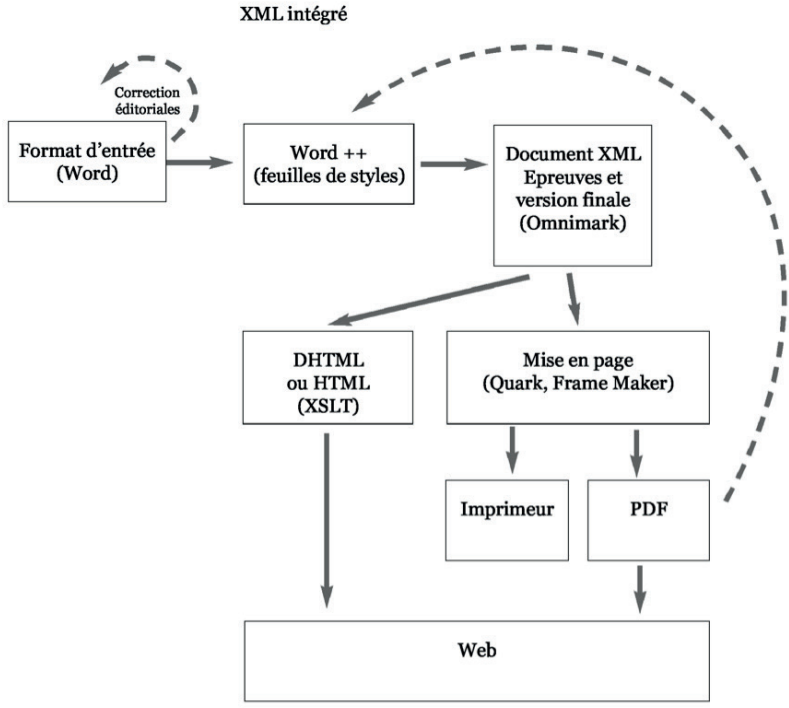
La norme XML n'est pas à proprement parler un format de document. Il s'agit plutôt d'un langage permettant de définir des formats de document. Ces formats sont des modèles de documents, que l'on appellera des schémas. Ces schémas expriment l'ensemble des contraintes que doit respecter un document XML.

Notons que, jusqu'à maintenant, les schémas ont toujours été représentés, autant en XML qu'en SGML, par des DTD, soit des définitions de types de documents. Toutefois, une norme, présentement en élaboration au W3C (XML Schema, voir http://www.w3.org/XML/Schema.html), permet également de définir un format XML. Cette future norme est appelée à remplacer les DTD, mais pour l'instant l'utilisation de celles-ci est recommandée, en particulier dans les applications documentaires où les XML Schemas apportent peu de nouvelles fonctionnalités intéressantes.

Les schémas et le respect de leurs contraintes constituent le cœur d'un système de gestion documentaire basé sur la norme XML. C'est pourquoi il faut leur accorder une grande importance lors de la conception et lors du développement des différents outils. C'est l'une des premières activités lors de la création d'un système XML. Les parties génériques du schéma devraient être inspirées de schémas de référence tels que Docbook, TEl ou ISO 12083. Normalement, un seul schéma devrait être suffisant pour un ensemble de revues, car les caractéristiques communes des articles sont plus nombreuses que leurs différences. En utilisant une approche générique pour certaines caractéristiques des documents, il est possible d'en arriver à un schéma unique. Toutefois, pour l'implantation de la chaîne de traitement pour plusieurs revues, il faudra évaluer l'intérêt d'utiliser un schéma de référence modulaire et de l'adapter pour obtenir un schéma spécifique pour chaque revue. En définissant correctement le schéma de référence et les schémas dérivés, il serait possible d'obtenir des outils uniques pour l'ensemble des revues et ainsi rendre les traitements efficaces. Le choix entre un seul schéma ou bien un schéma de référence plusieurs fois adapté doit être fait après une étude plus approfondie des caractéristiques propres des revues impliquées.

La production du document XML se fait automatiquement à partir du document Word. Pour y arriver, il est beaucoup plus facile de passer par le format RTF. La conversion d'un document RTF au format XML (respectant un schéma spécifique) n'est pas une tâche générique pour laquelle il existe des outils déjà préparés. Le XML n'étant pas un format, il est impossible de créer un outil de conversion du format RTF vers n'importe quel schéma XML.

Il existe deux types d'outil pour effectuer cette conversion. (1) La programmation pour tirer profit d'un langage pour manipuler les différentes structures du format RTF pour les convertir en XML. (2) D'autres solutions basées sur des règles (par exemple « les paragraphes qui sont en style Titre II sont des titres de section ») sont en général plus faciles à mettre en place, mais moins souples, moins puissantes.



L'utilisation d'un langage de programmation s'avère nécessaire si le contenu du document Word doit être manipulé de façon à changer l'ordre des éléments, à ajouter du contenu, à intégrer de l'information provenant de sources externes, etc. Une telle solution sera nécessaire dans le cas où, par exemple, les documents Word ne sont pas conçus en fonction du document XML à produire, mais plutôt en fonction d'autres impératifs, tels que les pratiques antérieures. Le langage de programmation Omnimark peut s'avérer une solution intéressante pour ce genre de conversion, tout comme le langage de transformation XSLT. Dans ce dernier cas, il faudra d'abord convertir le document Word vers un format XML quelconque, générique.

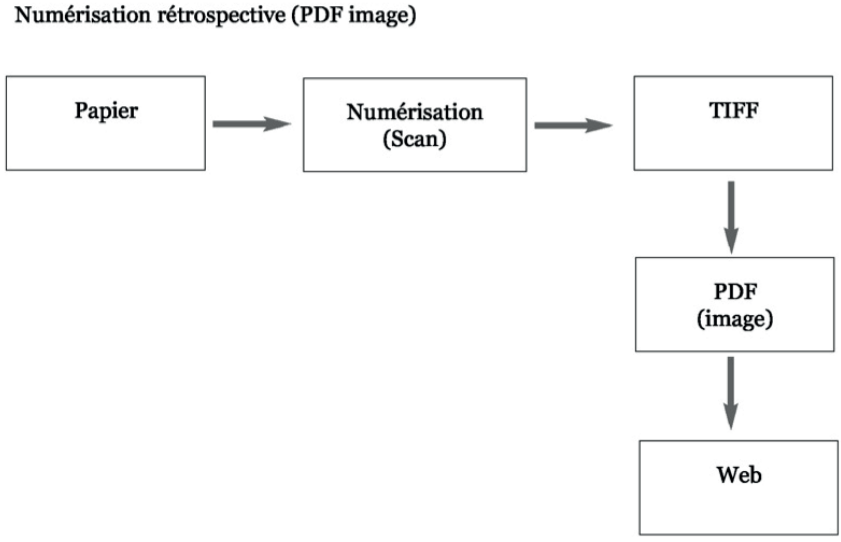
L'utilisation d'un outil basé sur des règles sera suffisante si les documents Word sont construits de façon similaire aux documents XML. Dans un tel cas, la structure peut être induite directement des styles utilisés, peu de contenu devra être déplacé et l'écriture de règles de structuration devient très efficace. Les outils les plus connus pour effectuer ces transformations sont Majix et UpCast.

Le choix final de l'approche et des outils doit tenir compte de critères comme les connaissances et l'expérience des personnes qui programmeront les conversions ou établiront les règles.

La numérisation rétrospective

La numérisation rétrospective consiste en la création d'un document électronique à partir d'un document, le plus souvent, sur support papier. Les articles de revues dont on possède la source électronique peuvent être convertis au format « Word ++ » et suivre le reste de la chaîne. Les articles dont on ne possède pas la source doivent être numérisés et traités par reconnaissance optique de caractères, ou encore re-saisis, puis amenés dans le format « Word ++ » et suivre le reste de la chaîne.

Cette dernière solution suppose de consentir de très grands efforts et implique des coûts importants, ce qui rend difficile la mise en route de grands projets de numérisation rétrospective et d'application d'une chaîne de traitement XML. À titre indicatif, gardons à l'esprit que la numérisation rétrospective en format image coûte dix fois moins cher la page que la numérisation + ROC. Il faut ensuite ajouter les coûts de traitement pour la vérification des fichiers (coquilles ROC) et la conversion vers le XML. Pour cette seule considération, on peut être porté à utiliser une approche différente, qui consiste à diffuser en mode image les articles déjà publiés, dans un format PDF par exemple, ce qui permet aux lecteurs de les consulter assez aisément. Cette solution temporaire doit être implantée en pensant à l'avenir, afin de rendre possibles de futurs traitements en reconnaissance optique de caractères, voire de conversion vers XML. En plus de respecter un niveau de qualité de la numérisation, le format image retenu doit assurer l'archivage à long terme, le format TIFF par exemple, et permettre la récupération des fichiers pour les traitements futurs.



Comme il est impossible d'effectuer une recherche plein texte dans un format image, des métadonnées doivent être associées aux articles numérisés pour assurer l'intégration des collections courante et rétrospective par l'outil de recherche et l'interface de navigation. L'ajout de ces métadonnées constitue toutefois une tâche manuelle qui peut être fastidieuse, même s'il est possible de récupérer une partie des métadonnées depuis une base de données bibliographiques.

La production des formats de diffusion de l'information

Les formats de diffusion sont tous produits à partir d'une seule source, soit le document XML de référence.

La consultation à l'écran

La consultation des articles à l'écran n'est pas toujours d'un grand confort et, comparativement à l'imprimé, la vitesse de lecture est réduite et la capacité de rétention de l'information est plus faible (Grabmeier J., 2000). Mais il reste que l'ordinateur fournit des outils très utiles et performants pour l'exploitation des informations disponibles dans le Web. Cela est

d'autant plus vrai lorsque l'environnement de consultation offre des aides à la lecture comme des annotations personnelles ou partagées ou encore des signets.

Le format de diffusion en ligne est indépendant de l'encodage initial du document en format XML. En effet, à partir d'articles de revues encodés en XML selon le même schéma, il est possible d'obtenir des présentations visuelles très variées pour les articles des différentes revues. Ainsi, il est facile de respecter les chartes graphiques des revues, et même de les mettre en valeur en utilisant les avantages de l'électronique.

Le format HTML — plus précisément le Dynamic HTML (DHTML) — constitue, encore pour quelques années au moins, le format de choix pour la diffusion d'information électronique. La conversion des documents XML vers un format DHTML se fait facilement et automatiquement, moyennant un effort initial de programmation.

Depuis novembre 1999, la norme XSLT, issue du W3C, est un langage de choix pour de telles transformations. Les processeurs XSLT sont particulièrement efficaces pour des documents de taille petite ou moyenne, ce qui est tout à fait approprié pour les articles de revues. Cette norme facilite, de plus, la création de feuilles de styles modulaires qui permettent de créer plusieurs variations d'une même transformation. Ce dernier aspect est important, car pour offrir un grand confort de navigation aux utilisateurs, il y a lieu de préconiser la production de plus d'une version DHTML des articles. De façon à respecter l'image graphique de chaque revue, cette partie de la chaîne de traitement implique la création d'un grand nombre de programmes de conversion légèrement différents, d'où l'importance d'une approche modulaire, ce que permet la norme XSLT.

Les conversions au format HTML peuvent être effectuées de façon dynamique, c'est-à-dire à chaque fois qu'un article est demandé sur le serveur de diffusion.

Elles peuvent aussi être faites à l'avance et déposées sur le serveur. La première approche minimise les besoins en espace disque, mais demande une plus grande puissance du serveur. La deuxième approche

inverse ces besoins. Puisque les articles, une fois publiés, ne sont pas modifiés, et que le coût de l'espace disque est peu élevé, le pré-traitement des articles semble l'avenue la plus appropriée. Toutefois, si on veut offrir aux utilisateurs une représentation HTML qui dépend de certains paramètres inconnus à l'avance, une approche dynamique est nécessaire. Par exemple, si le serveur offre des outils d'annotation publique et privée, la présentation des documents dépendra des annotations et des préférences d'affichage, ce qui exclut le pré-traitement des articles.

La diffusion des documents XML directement dans le Web se fait de deux façons. D'abord, on peut associer à ces documents une feuille de style CSS. Cette approche est peu recommandée en raison de ses limites : les feuilles de styles CSS permettent de créer une mise en page correcte d'un document XML, mais offrent peu d'intérêt en matière de dynamisme des pages ou de mise en contexte de l'information dans l'ensemble d'un site. L'autre approche consiste à associer une feuille de style XSLT au document et, dans ce cas, la transformation s'effectue sur le poste client. Pour l'instant, seul le navigateur Internet Explorer offre cette fonction. À moyen terme, il s'agit d'une stratégie intéressante, pour minimiser à la fois les traitements du côté du serveur et l'espace disque requis.

L'impression à distance

L'impression à distance consiste à fournir aux lecteurs, habituellement via Internet, un document mis en page et destiné à être imprimé sur leur imprimante personnelle. Cette approche évite la lourdeur de l'impression professionnelle et répond à une diversité de besoins. De plus, les lecteurs peuvent imprimer seulement les pages ou les articles dont ils ont besoin, ce qui évite une consommation démesurée de ressources.

Les logiciels de navigation Web permettent d'imprimer les documents consultés. Étant conçus pour la consultation en ligne, la mise en page de ces documents (le plus souvent HTML) n'est pas de bonne qualité, car les documents n'ont pas nécessairement été préparés pour l'impression. Le meilleur format d'impression à distance reste le format PDF d'Adobe. La production de ces documents implique les mêmes

opérations que la production des documents Postscript. Puisque l'impression professionnelle est encore d'actualité, il est tout à fait logique d'utiliser le produit intermédiaire de l'impression professionnelle, soit le document Postscript, pour obtenir un document PDF. Cette conversion est automatique et ne pose pas de problème particulier.

Par conséquent, la production des versions pour l'impression professionnelle et pour l'impression à distance est semblable ; seule une étape automatique s'ajoute pour obtenir des documents PDF à partir des documents Postscript. Notons au passage, que lors de cette conversion, des ajustements peuvent être effectués pour, notamment, réduire la qualité des images, évitant ainsi de trop grands temps de téléchargement des articles.

L'impression professionnelle

L'impression professionnelle permet d'obtenir des articles sur support papier, avec une très bonne qualité d'impression et habituellement reliés en numéros. Cette impression demande une expertise et des équipements particuliers. Cette tâche est habituellement effectuée par des sociétés spécialisées : l'imprimeur reçoit les documents mis en page, le plus souvent en format Postscript.

La technique la plus intéressante pour produire des documents Postscript à partir des documents XML est d'utiliser un logiciel de mise en page. Il s'agit d'importer l'information (les articles) dans un tel logiciel et de produire un document Postscript à la suite de la mise en page. En raison de la complexité de la mise en page pour l'impression, cette opération peut difficilement se réaliser de façon purement automatique, mais on peut s'en approcher. Par exemple, les logiciels Quark XPress et FrameMaker importent de l'information dans un langage balisé (XPress Tags pour XPress, MIF pour FrameMaker) que l'on peut obtenir assez facilement par conversion d'un document XML. Il existe également des solutions entièrement automatisées, avec par exemple 3B2 ou Adept Publisher, mais les efforts de paramétrage et de programmation sont très grands. Il existe une version de FrameMaker (FrameMaker + SGML) qui importe directement des documents SGML, mais cette option n'est pas particulièrement intéressante pour

notre chaîne de traitement basée sur XML, entre autres parce qu'il faut de toute façon manipuler le XML pour le convertir en SGML et l'adapter aux exigences de l'impression.

Il est difficile de produire un document entièrement imprimé de façon automatique, sans vérification manuelle, mais la chaîne de traitement proposée permet de minimiser ces interventions humaines, si bien que les premières épreuves pour l'impression se font presque automatiquement, avec seulement quelques petits ajustements manuels. Éventuellement, il sera possible de créer les épreuves papier avec la version XML et une feuille de styles CSS appropriée, mais, pour l'instant, le support CSS est trop peu développé dans les logiciels pour utiliser cette approche.

Les métadonnées

On ne peut prétendre diffuser une collection d'articles simplement en la rendant accessible dans le Web. C'est l'équivalent d'ouvrir les portes d'une immense bibliothèque sans fournir de catalogue : on se doute bien qu'elle contient des informations intéressantes, mais aurions-nous le temps de les chercher dans ces conditions ? Le Web ressemble actuellement à cette « bibliothèque ». Les métadonnées sont un des outils pour assurer la diffusion efficace et optimale des articles. L'édition savante électronique transforme le processus traditionnel, notamment pour le repérage. L'exploitation des métadonnées appliquées à des articles de revues en version électronique permet de tirer de grands avantages dans le repérage et la mise en réseau de sources documentaires diverses.

Les métadonnées : raison d'être et usages

L'analogie classique pour décrire les métadonnées, c'est la fiche de carton d'un catalogue de bibliothèque. Cette petite fiche permettait de faire un choix sans pour autant voir le document décrit. Les informations bibliographiques, les résumés, les termes d'indexation, les abstracts, tout ce qui peut être un substitut au document original et qui libère les usagers potentiels de la nécessité de connaître à l'avance l'existence et les caractéristiques de ce document sont des métadonnnées.

Dans l'univers électronique, les métadonnées sont des informations à propos d'objets numériques qui peuvent être soit des articles, soit d'autres objets numériques insérés dans ces articles (par exemple des fichiers image ou son). Les métadonnées décrivent les attributs et le contenu de ces objets. Elles sont utiles au repérage, mais également à la gestion, à la description, à l'accès et à la conservation de l'information.

L'utilisation des métadonnées est nécessaire par une simple raison de logique mathématique. En juin 1993, on comptait 130 sites Web (Gray M.), en novembre 2000, la société Netcraft en comptait 23,8 millions (Netcraft). Les chiffres sont encore plus impressionnants si on considère les documents électroniques présents sur ces sites : quelque Boo millions de documents publiquement accessibles par les robots de recherche dans le Web étaient recensés en 1999 (Lawrence Steve C. et Giles Lee, 1999). Michael Dahn (2000), sans contester l'enquête de Lawrence et Giles, en vient à la conclusion qu'en novembre 1999 le Web publiquement indexable comprend 1,16 milliard de documents, mais que le Web publiquement accessible se chiffre plutôt entre 1,45 et 2,33 milliards de documents (voir plus loin le commentaire concernant Web accessible/Web indexable). De leur côté, les producteurs du moteur de recherche Inktomi ont atteint leur milliardième document le 18 janvier 2000 (Inktomi). Plusieurs motifs plaident pour le recours à des métadonnées dans cet univers à la fois riche et pléthorique.

Tout d'abord, l'accessibilité et l'utilisation accrues des documents électroniques, grâce notamment aux facilités de recherche, doivent être supportées de façon conséquente par les outils offerts aux utilisateurs, d'où, au premier chef, les métadonnées. Elles améliorent aussi grandement la recherche d'information dans de multiples collections et permettent, par exemple, l'interopérabilité entre différents portails. Il s'agit, de plus, d'un outil précieux pour diversifier les points d'accès à l'information, la présentation des résultats et les possibilités de manipulation de l'information. Les métadonnées sont également des outils de gestion de la protection des droits et des restrictions de consultation.

On estime que seulement 6 % du Web est à caractère scientifique ou éducatif ; la plus grande masse du contenu (83 %) étant à caractère commercial (Lawrence S.C. et Lee G., 1999). Dans ces conditions, le simple fait de rendre un texte savant disponible dans le Web équivaut aujourd'hui à verser un verre d'eau de plus dans l'océan.

De là, deux tendances lourdes se dégagent pour mettre un peu d'ordre dans ce chaos : la création de portails spécialisés et l'utilisation de métadonnées. On assiste de plus en plus à une spécialisation du Web. Bientôt, on retrouvera des Webs à l'intérieur du Web. Une de ces manifestations est la création de portails : portails de divertissement, portails de services financiers, portails de revues savantes, etc. Lorsque les portails scientifiques et l'infrastructure de navigation et d'interopérabilité entre ces portails seront implantés et consolidés, les chercheurs n'auront plus à naviguer sur tout l'océan Web ; on s'orientera vers la qualité plutôt que vers la quantité. À cet égard, les métadonnées sont le moyen pour assurer cette mise en ordre du Web.

Les métadonnées : les créer, les stocker

Dans un environnement de documents structurés, un certain nombre de métadonnées sont déjà présentes grâce à la structure des documents. Par exemple, le titre d'un article est déjà identifié comme tel par les balises XML. Il est possible d'extraire ces éléments d'information du contenu du texte pour générer des schémas normalisés de métadonnées, comme le Dublin Core. Tous les éléments de description bibliographique de l'article, de même que différents types d'indexation matière, devraient se retrouver dans les métadonnées.

Le principe à suivre est d'associer les métadonnées le plus tôt possible, car elles peuvent être utiles dans le processus de production. Cela dit, en raison de la rigueur requise pour cette opération, il est souhaitable de confier aux éditeurs de revues savantes plutôt qu'aux auteurs la création des métadonnées bibliographiques dans le document « Word ++ », et ce, au tout début du processus de mise en forme du document reçu des auteurs.

Il est facile d'associer des métadonnées, et de façon très structurée, aux documents XML. Toute métadonnée propre à l'article devrait s'y

retrouver en premier lieu, par insertion, quitte à ce qu'elle soit reprise ailleurs dans le système, par souci d'efficacité.

Pour les articles rétrospectifs, les métadonnées devraient être associées au moment de la numérisation, à l'aide d'un traitement manuel ou encore par récupération semi-automatique des informations depuis des bases de données bibliographiques.

Quelques modèles de métadonnées

Les éditeurs utilisent, de plus en plus, un standard d'identification unique des articles, le DOI (Digital Object Identifier; voir http://www.doi.org/). Cet identificateur permet, entre autres, de donner des adresses permanentes aux articles, ce qui facilite leur repérage et leur gestion. L'ajout d'un identificateur unique aux articles ouvre la porte à de nombreuses applications, y compris une navigation facilitée dans les références bibliographiques. D'autres outils proposent des solutions équivalentes, entre autres, le PURL (http://www.purl.org) et SFX (Van de Sompel, 1999).

En apportant une réponse au problème des liens URL rompus (erreur 404), ces services implantent électroniquement, et de façon pérenne, cette caractéristique fondamentale de l'édition savante : les citations. Le service CrossRef (http://www.crossref.org/faqs.htm) a pour objectif de lier les références bibliographiques aux contenus des articles diffusés en ligne. On estime qu'à la fin de l'an 2000, trois millions d'articles provenant de milliers de périodiques auront été liés au moyen de CrossRef et que la croissance annuelle sera de 500,000 articles. Notons que les liens ne seront pas seulement entre des revues savantes, mais pourront pointer vers des articles d'encyclopédie, des actes de colloque, des manuels scolaires, etc., ce qui permettra un enrichissement important de la lecture. CrossRef s'appuie fortement sur les DOI.

Dublin Core (DC; Haigh S., 1999) est le standard de métadonnées le plus répandu et le plus avancé pour la description des ressources Internet. Créé en 1995 à Dublin en Ohio (siège de Online Computer Library Center), le développement du Dublin Core est assuré par le Dublin Core Directorate, supervisé par le OCLC Office of Research

and Special Projects (Morgan C., 1999, p. 192 ; Hudgins J. et al., 1999, p.14).

Le Dublin Core comprend 15 éléments de base pour décrire les ressources électroniques. Il est conçu autour de 5 principes fondamentaux, à savoir :

- Tous les éléments sont optionnels.
- Tous les éléments sont répétables.
- DC est extensible. DC est un plus petit dénominateur commun (DC Simple), mais permet aussi, si on le désire, d'avoir une description plus riche, au moyen de sous-éléments (DC Qualified).
- DC est multidisciplinaire.
- DC est international (plus de 20 langues actuellement).

Ce modèle de métadonnées est déjà utilisé dans le milieu de l'édition savante. L'éditeur John Wiley & Sons, qui publie chaque année 35,000 articles dans plus de 400 revues, en a fait son standard. Ces revues sont codées en SGML. L'en-tête (*header material*) comprend des informations sur la revue, le volume et le numéro, mais on y retrouve aussi le titre de l'article, l'auteur et son affiliation, le résumé, des mots-clés, la date de réception de l'article, etc. (Morgan C., 1999, 194).

Une autre norme importante augmente les possibilités d'exploitation des métadonnées. RDF (Resource Description Framework) est une norme pour faciliter le traitement des métadonnées. Il fournit l'intéropérabilité entre les applications qui échangent de l'information non compréhensible par les machines du Web (http://www.la-grange.net/w3c/REC-rdf-synta/). En plus d'être utilisé pour la découverte de ressources, le catalogage, l'évaluation du contenu, la gestion des droits d'auteur, il sera également possible avec RDF d'insérer des métadonnées qui informeront les lecteurs des pratiques en matière de protection des renseignements personnels grâce au projet « Platform for Privacy Preferences » (P3P ; http://www.w3.org/P3P/Overview.html/). Par exemple, dans une transaction électronique, il sera possible d'informer les clients que leurs informations nominatives ne seront pas transmises à des tiers. Ces échanges d'information se feront auto-

matiquement et de façon transparente entre le navigateur du client et le serveur Web. Cette norme est essentielle au développement de la confiance en matière de commerce électronique (« Web of trust »).

LA DIFFUSION

Le Web est sans conteste un média privilégié par la communauté scientifique qui cherche à maximiser la communication des connaissances. Pourtant, la multitude d'informations disponibles, le manque d'outils de classification et les déficiences dans leur organisation constituent des entraves à l'accès au savoir. Une simple présence des revues en sciences sociales dans Internet, serait-ce par un portail unique, ne garantit pas une visibilité et un rayonnement adéquats à l'information scientifique véhiculée par ces médias. Encore une fois : être disponible sur le Web n'est pas synonyme d'être diffusé.

La diffusion est une composante essentielle, voire stratégique, de l'édition électronique. Il est judicieux d'élaborer des stratégies de diffusion et de mise en place adaptées à cet environnement virtuel et d'y allouer les ressources humaines et financières suffisantes. Les différents aspects qui participent au processus de diffusion doivent être distingués (accessibilité et ergonomie du site, formats de diffusion et services offerts, actions stratégiques de positionnement dans le Web), ce qui permet d'en montrer les enjeux et les retombées éventuelles.

Une interface Web pour diffusion des revues savantes

L'optimisation de la diffusion d'une revue ou d'un bouquet de revues est tributaire de plusieurs éléments dont, au point de départ, l'attention accordée à la conception du site Web (Lynch P.J. et Horton S., 1997). L'accessibilité est l'une des caractéristiques permettant au site de devenir un outil de communication et un foyer de référence pour la recherche. Cette accessibilité dépend de (1) l'environnement de communication de l'information et (2) de l'aisance avec laquelle on peut mener la consultation de l'information.

L'accès à l'information est façonné par deux grands paramètres. L'architecture d'un site désigne la structuration de l'arborescence

des fichiers et l'organisation de l'information. L'architecture devrait varier les modes d'accès à l'information, permettre l'exploitation de la puissance du lien hypertexte et tirer profit des fonctions offertes par le site. Précisément, l'accès est supporté par les infrastructures de repérage de l'information, d'où l'intérêt de diversifier les modes d'accès au contenu ; on pense, par exemple, à l'accès par catégories de répertoire, par un module de recherche plein texte et/ou par champs (titre, auteur, description, mots-clés, etc.), ainsi que par une classification documentaire.

La consultation de l'information sera facilitée par le soin apporté à l'ergonomie de l'interface et par la puissance des outils d'aide à la navigation. L'ergonomie commence par des précautions toutes simples. Le graphisme, qui confère au site son identité visuelle, devrait viser la sobriété et la valorisation du contenu informationnel. Un environnement efficace se caractérise par la simplicité, la rapidité et la clarté. Des moyens assez simples y contribuent, tels qu'un plan du site, des indications de positionnement pour que l'utilisateur se situe dans l'architecture du site, et des fichiers d'aide se rapportant aux modes d'accès à l'information.

L'accessibilité au contenu informationnel et les conditions de consultation sont au cœur de la conception d'un site puisque ces deux dimensions contribuent à assurer une large diffusion. Ces dimensions valorisent la richesse du contenu et sa mise en forme, qui constituent les éléments de pertinence d'un site. Mais le contenu ne se suffit pas à lui-même, encore faut-il s'assurer que le site ait une visibilité maximale auprès des utilisateurs potentiels, notamment en veillant à ce qu'il soit répertorié par les outils de recherche.

Le système de diffusion

Le système de diffusion d'un site combine plusieurs services dont la mise en route peut s'étaler dans le temps en fonction des priorités et des ressources disponibles. Dans cette perspective, il est utile de qualifier les divers services pour établir un programme d'actions. C'est en ce sens que, dans l'aperçu qui suit, les services ont été regroupés en services de base et en services complémentaires.

Services de base

On peut dénombrer au moins quatre services de base que devrait réunir un site de diffusion de revues.

Le lecteur doit pouvoir naviguer aisément et toujours comprendre ce qu'il peut trouver en poursuivant sa route. Le premier type de navigation suit l'organisation logique des périodiques scientifiques, telles les relations entre une revue, ses volumes, ses numéros et ses articles. Outre les liens à créer entre les différentes unités d'information, un mécanisme de fiches descriptives permet de prendre rapidement connaissance du contenu d'une unité d'information. Pour une revue, la fiche descriptive comprend le nom, la périodicité, les objectifs, les personnes responsables, etc. Pour un numéro, la table des matières remplit cette fonction, alors que, pour un article, le titre, l'auteur ou les auteurs, le résumé apparaissent naturellement sur une telle fiche. Le défi du système de navigation consiste à modéliser correctement les différentes fiches descriptives et à les produire automatiquement depuis les unités d'information fondamentales que sont les articles de revue. Certaines informations ne pourront être extraites des articles et devront être produites à part, mais ces interventions devraient être minimisées en étant intégrées dans un processus automatique lors de la production.

Une navigation thématique devrait être offerte. Il s'agit d'exploiter les métadonnées associées aux articles pour offrir une navigation non plus basée sur l'organisation par revues ou numéros, mais sur des concepts comme des sujets, des auteurs ou des relations thématiques. L'analyse des métadonnées permet d'identifier celles qui pourront donner lieu à un système de navigation.

La consultation devrait débuter par la fiche descriptive d'un article. La fiche descriptive devient le point d'entrée de l'article dans la mesure où la diffusion d'un article peut se faire selon plusieurs formats et représentations, et qu'aucun de ces formats n'est, a priori, prioritaire. D'ailleurs, l'adresse d'un article devrait être reliée à celle de sa fiche descriptive. Par exemple, dans un système basé sur les DOI, la résolution d'un DOI devrait mener le lecteur à la fiche descriptive de l'article en question. Cette fiche descriptive doit contenir un lien

vers toutes les versions disponibles de l'article ; une fois ce lien activé, le lecteur peut consulter l'article à sa guise.

La recherche d'informations est essentielle dans un site, en particulier lorsqu'il réunit des revues savantes. Lorsque la collection de documents est assez volumineuse, la navigation ne suffit pas. L'outil de recherche proposé doit permettre de distinguer l'information primaire, c'est-à-dire les articles scientifiques, l'information secondaire, par exemple la description d'un numéro ou d'une revue, et l'information de navigation (qui devrait normalement être ignorée en recherche).

Un bon outil de recherche permet (1) de composer des requêtes de recherche variées et puissantes et (2) de présenter les résultats de façon claire et utile pour les utilisateurs.

1. Pour répondre au premier objectif, les documents indexés sont les versions XML de référence des articles. L'outil de recherche doit pouvoir tenir compte non seulement du texte des articles, mais aussi de leurs métadonnées ainsi que de la structure des documents. Les métadonnées et la structure sont conçues pour pouvoir préciser les critères de recherche — par exemple, de chercher des mots seulement dans les sujets ou les auteurs ou les titres de sections —, mais aussi pour influencer le tri par ordre de pertinence des résultats.
2. Le deuxième objectif, relatif à la présentation des résultats, est plus complexe. A priori, la possibilité d'effectuer différents tris, y compris par ordre de pertinence, ne semble pas poser de problèmes fondamentaux. Toutefois, une fonction intéressante des outils de recherche consiste à mettre en évidence les mots cherchés dans les documents trouvés. Puisque la recherche s'effectue dans les documents XML, en utilisant notamment leur structure, mais que les documents peuvent être consultés dans les formats HTML ou PDF, il n'est pas facile d'établir une correspondance entre la requête présentée et le contenu des articles dans ces formats. Des développements s'imposent de ce côté et il convient de les suivre et d'intégrer de telles fonctions à plus long terme.

La diffusion libre et gratuite des articles de revue dans Internet ne nécessite pas la gestion des utilisateurs et des droits d'accès. Cette gestion devient nécessaire dès que des restrictions sont imposées pour l'accès aux documents ou dès que des services personnalisés sont offerts aux utilisateurs.

Un système de gestion des utilisateurs suppose l'élaboration d'une politique d'accès à l'information, l'identification des unités d'information et des services, qui peuvent faire l'objet de restrictions, et la définition de ce qu'est un utilisateur. Une fois ces éléments établis, un système qui permet de combiner tous ces concepts et toutes ces définitions, doit être mis en place. La plupart des systèmes de gestion documentaire offrent une gestion des utilisateurs par groupes et personnes, de même qu'une gestion des droits d'accès par document ou par dossier. À l'aide de ces outils, on peut arriver à offrir une gestion qui remplit une bonne partie des besoins exprimés.

Services complémentaires

À ces services centraux, d'autres peuvent éventuellement s'ajouter pour valoriser la consultation et l'exploitation de la documentation diffusée.

La dissémination sélective de l'information consiste à envoyer par courriel de l'information directement aux personnes concernées. En raison de la périodicité des revues savantes, il s'agit d'un service très pertinent pour répondre aux besoins des chercheurs. Pour être efficace, un système de dissémination sélective de l'information doit être autogéré par les utilisateurs, de leur identification à leurs préférences, en plus de leurs besoins en information. Comme pour la navigation, la dissémination devrait se faire selon deux orientations, la première, basée sur la publication des articles, la deuxième, sur les préférences thématiques. Dans le premier cas, les utilisateurs sont avertis de la publication d'un article, et donc habituellement d'un numéro d'une revue, s'ils se sont déclarés intéressés par cette revue. Dans le deuxième cas, les utilisateurs sont saisis de la parution d'un article lorsqu'il correspond à certains sujets ou certains auteurs préalablement identifiés. L'aide à la lecture désigne les fonctionnalités qu'il est possible d'ajouter afin de faciliter le travail du lecteur. Rien

n'oblige d'implanter ces fonctionnalités dès la première phase de la diffusion des revues, bien que, pour certains utilisateurs, il s'agira de raisons suffisantes pour apprécier le système mis en place.

On peut s'inspirer des fonctionnalités d'aide à la navigation qu'offrent différents logiciels de navigation de contenu documentaire.

1. L'annotation est une fonction intéressante, car les lecteurs ont souvent l'habitude de prendre des notes lors de la consultation d'articles scientifiques. Cette fonction est encore plus avantageuse lorsque les annotations peuvent être partagées entre différents utilisateurs.
2. Les signets permettent de revenir rapidement à certains éléments d'information. De plus, pour des articles imposants, la possibilité de marquer un endroit précis à l'intérieur du document peut s'avérer indispensable.
3. Les liens créés par l'utilisateur peuvent aussi être un outil intéressant d'acquisition de connaissance dans un corpus documentaire, en particulier lorsque ces liens peuvent être partagés ou « publiés » afin de les faire connaître à d'autres personnes.

Ces trois fonctionnalités peuvent être implantées à l'aide d'outils respectant la norme XLink, qui permet de créer des relations entre des documents ou des parties de documents, sans nécessairement toucher à leur contenu. Le partage des annotations, signets et liens implique l'implantation de ces outils « côté serveur ».

4. La manipulation des images et des autres contenus multimédias est une autre fonctionnalité intéressante. Pour l'utilisateur, il peut être très utile de modifier les dimensions d'une image pour mieux apprécier les détails, de démarrer et arrêter une séquence vidéo, etc. Pour y arriver, des logiciels « côté client » doivent être fournis, car les navigateurs Web actuels n'ont pas ces fonctionnalités.

La personnalisation d'un service Web consiste à offrir à l'utilisateur une expérience de navigation et de consultation adaptée à ses besoins, potentiellement différents de ceux d'un autre utilisateur. La personnalisation la plus intéressante pour un site de revues (encore

faut-il en réunir un certain nombre pour que la chose prenne un sens) consisterait à recueillir les préférences des utilisateurs en matière de thématiques, d'auteurs, de revues, etc., et d'offrir à ces utilisateurs, en première ligne, des articles qui respectent ces préférences. Une telle personnalisation n'est pas essentielle, mais représenterait une valeur ajoutée justifiant ainsi son implantation éventuelle.

Une offensive de diffusion sur la toile

Le référencement et le positionnement stratégique d'un site dans les différents outils de recherche Internet (index et répertoires) s'avèrent d'une importance cruciale pour attirer les utilisateurs (Stanley T., 1997a, 1997 b, 1997c. ; Liberatore K., 2000 ; Murphy K., 1996 ; Georgia Institute of Technology, 1998-10/). Mais la diffusion du contenu de revues savantes nécessite également une visibilité dans les principales bases de données disponibles au chercheur d'information dans le domaine des sciences sociales. D'où l'intérêt d'identifier ces bases de données potentielles et leurs revendeurs, d'examiner la représentation des revues dans celles-ci, d'analyser les possibilités offertes par ces bases et d'établir les principes d'une stratégie pour accroître leur dépouillement au sein de ces outils. Finalement, pour attirer de nouveaux publics, fidéliser les utilisateurs actuels et adapter les services offerts par le site aux nouvelles tendances, une campagne de webmarketing peut être envisagée.

Le référencement et le positionnement stratégique

Les internautes utilisent massivement (85 %) les outils de recherche pour trouver de l'information dans le Web et, dans une proportion encore plus grande, ils ne prennent pas connaissance de plus de deux ou trois pages des résultats de leur requête de recherche (Georgia Institute of Technology, 1998-10/). On constate l'importance, pour un site de revues savantes en sciences humaines et sociales, d'apparaître dans les bases de données des principaux outils de recherche du Web — c'est ce que l'on appelle le référencement — et de sortir dans une position privilégiée dans les résultats des requêtes acheminées par les utilisateurs à ces moteurs — ce qui renvoie à la notion de positionnement stratégique. Le référencement

et le positionnement sont étroitement liés et font partie intégrante des objectifs de diffusion d'un portail.

Le référencement est un processus circulaire et évolutif : l'inscription n'est jamais définitive et doit être adaptée aux nouveaux besoins de visibilité. Il faut y consacrer des ressources, mais surtout considérer cet aspect au moment même de la conception du site Web. Une action fructueuse demande une connaissance à la fois des outils de recherche, des modalités d'indexation et des politiques éditoriales des répertoires. Ici on doit distinguer (1) les outils de recherche (2) des répertoires ou annuaires.

Pour les outils de recherche, ce sont des robots qui indexent de façon automatisée un site ayant ou non soumis une demande pour être intégré à la base de données. La connaissance des modalités d'indexation permet d'optimiser les pages Web pour les divers outils (pensons à : AltaVista, Google, Northern Light, America Online Search, FAST Search, Netscape Search, etc.).

Pour les répertoires ou annuaires, qui classifient les ressources Internet à la suite de l'évaluation « d'éditeurs » humains et en référence à une politique de classement, la démarche est différente. Les « éditeurs » humains décident, en fonction de critères, tels que la qualité du contenu, l'organisation de l'information, l'accessibilité au contenu, la fréquence de mise à jour, la pérennité de l'information, etc., d'intégrer ou non le site au répertoire (à titre d'exemple, citons : Yahoo!, La toile du Québec, Lycos, Looksmart, Francité, Nomade, Open Directory).

À défaut d'être partout, il est préférable de canaliser les ressources et l'attention pour le référencement vers les outils de recherche majeurs, qui sont les plus susceptibles d'attirer beaucoup d'utilisateurs. Cela, évidemment, sans négliger l'intérêt d'apparaître dans des outils de recherche spécialisés qui s'adressent à des populations plus ciblées, mais qui correspondent à la mission du site. À ce stade, seulement une fraction du chemin est parcourue, car une bonne visibilité dépend de la capacité à figurer dans les premiers résultats d'une requête de recherche, ce qui conduit à parler du positionnement stratégique.

Le positionnement fait référence à un ensemble de méthodes employées au cours de l'élaboration, de la publication et de l'entretien du site du portail pour assurer une visibilité optimale dans les outils de recherche et les répertoires. Le positionnement sera qualifié d'optimal si le site du portail apparaît dans les 20 ou 30 premiers résultats à la suite d'une requête dans un outil de recherche. Pour les répertoires, l'objectif, c'est d'être intégré dans la base de données. À cette fin, il convient de recenser, pour s'y adapter, les critères utilisés par les outils de recherche des répertoires.

L'étude des mots-clés et des catégories par lesquels le site devrait être référencé, ainsi que des rubriques sous lesquelles le site devrait apparaître dans les répertoires ou annuaires, est capitale pour assurer le meilleur positionnement, mais aussi pour avoir une bonne correspondance avec le vocabulaire que l'utilisateur potentiel est susceptible d'utiliser dans sa recherche.

Une surveillance régulière est de mise pour s'assurer du positionnement du site dans les outils de recherche, car les algorithmes de calcul de pertinence des outils changent, ce qui demande de s'adapter en conséquence. Concurremment, de nouveaux sites apparaissent et désirent occuper les mêmes créneaux. Il est en ce sens parfois nécessaire d'ajuster la démarche en référence/positionnement.

La visibilité dans les bases de données

En plus de la visibilité atteinte par les outils de recherche et les répertoires, il faut se préoccuper de l'accessibilité et du repérage du contenu informationnel du portail dans les bases de données importantes qui dépouillent le contenu des périodiques en sciences humaines et sociales. Ce n'est pas en soi une nouveauté : les directions de revue prennent déjà plusieurs initiatives pour s'assurer que leurs publications imprimées soient indexées par plusieurs bases de données, en se concentrant surtout sur les plus prestigieuses ou les mieux placées selon le domaine de spécialisation.

Ces bases de données se veulent sélectives au sens où elles retiennent les revues en fonction de critères de qualité et de rayonnement dans le domaine de spécialisation. La langue joue aussi un rôle : très peu de bases de données indexent massivement des documents en français —

très peu de bases proviennent d'ailleurs du monde francophone — et les bases de données de langue anglaise sont assez sélectives, notamment sous l'angle de la langue de communication.

L'objectif consiste à accroître la visibilité des revues en sciences humaines et sociales, tant dans les bases de données (les index) francophones qu'anglophones. Pour se ménager un meilleur accueil dans ces dernières, il est judicieux de produire des résumés « consistants « en anglais pour les articles de chaque revue diffusée sur le site. D'ailleurs, ces résumés, plus formalisés que ceux généralement proposés, auraient un triple avantage supplémentaire : 1° améliorer les conditions de sélection dans les grandes bases de données ; 2° faciliter la consultation sur le site et la prise en compte d'articles par des utilisateurs non francophones ; 3° enrichir les métadonnées (qui puisent dans les résumés). Ces résumés (Hartley J., 1997, p.313-317) pourraient comprendre certaines rubriques, telles que 1° contextualisation, 2° objectif, 3° méthodologie, 4° résultats, et 5° conclusion. Il s'agit d'un outil de communication scientifique dont l'utilité serait certaine, entre autres pour l'indexation dans les bases de données.

Les outils de promotion pour accroître la visibilité du site

Une offensive de diffusion sur la Toile passe aussi par une préoccupation touchant la promotion du site lui-même. L'accroissement de la visibilité, l'augmentation du nombre d'utilisateurs et l'impact du portail dans et sur la communauté scientifique internationale sont fonction de la capacité de faire du portail un lieu dynamique intellectuellement, qui nourrit, accueille et provoque des activités en mesure d'interpeller le milieu des chercheurs nationaux et internationaux.

La communication et l'interaction du site avec les chercheurs et les étudiants constituent un élément de première importance. Pensons aux actions suivantes : offrir des infrastructures pour recueillir les commentaires des utilisateurs, animer une liste de discussion ou un forum favorisant l'échange d'idées scientifiques, organiser des événements sur le site (débats, conférences virtuelles, diffusion en primeur de certains contenus, entrevues en direct, etc.), publier une lettre d'information

à périodicité fixe par courriel, alimenter une section « actualités » à l'intérieur du site, fournir l'accès libre au maximum d'information.

La pertinence, la qualité et l'intérêt des services parlent d'eux-mêmes. Leur conception et leur mise en application doivent prendre en compte ces dimensions. Parmi ceux-ci, relevons : un service de diffusion sélective de l'information, la personnalisation de l'interface avec le portail et un service d'aide à la lecture, la mise à disposition d'une liste de liens pertinents vers des sites d'intérêts connexes. Le site, comme acteur dans la communication scientifique, a tout avantage à s'inscrire dans une logique de réseau à l'échelle internationale, d'où l'intérêt stratégique des partenariats. Certaines pratiques simples y concourent : établir un réseau de sites de revues savantes ; procéder à des échanges de liens réciproques avec des sites poursuivant les mêmes missions ; mettre en place un programme d'échange de bannières publicitaires (forme de publicité complètement gratuite et ciblée) ; parrainer des sites de moindre envergure traitant des mêmes sujets afin de les aider à obtenir de la visibilité tout en favorisant l'échange de services.

Le recours à la publicité aurait d'abord pour objectif de faire connaître les services du site. Il s'agit, pour l'essentiel, d'une démarche d'information. Pensons à : des envois massifs aux groupes de discussions et aux listes de distribution pour « publiciser » le site-portail, à des envois massifs par courriel à des usagers potentiellement intéressés par les services offerts par le site (prospection), à un communiqué de presse annonçant le site aux journalistes susceptibles de relayer l'information dans leurs publications respectives (médias traditionnels et nouveaux médias), à une conférence virtuelle avec des journalistes du domaine et des acteurs importants du milieu pour faire connaître les services offerts par le portail.

La promotion devrait miser sur le dynamisme et la créativité du site. Il n'en reste pas moins que le facteur promotionnel le plus crucial pour un site de revues savantes, c'est sa richesse documentaire, ses services, la qualité et le renouvellement de son information.

L'ARCHIVAGE

La conservation des revues savantes est assurée par différents organismes bibliothèques, centres d'archives et éditeurs, suivant des procédures et méthodes, polies par les ans. Si toutes les règles de l'art sont suivies, on peut affirmer sans crainte qu'un document papier sera encore lisible dans 500 ans. Il est difficile de faire une telle affirmation pour les documents numériques. Tout comme on l'a fait pour la préservation du papier, les procédures permettant d'assurer la pérennité de l'accès aux documents numériques sont en train de se définir. Dans le cas des revues savantes, il s'agit là d'une question fondamentale. D'abord, parce que les revues savantes représentent un investissement important en temps et en argent. Ensuite et surtout, parce que ces revues colligent l'histoire et l'évolution des disciplines et qu'elles font partie de notre patrimoine intellectuel et culturel.

Plusieurs aspects sont à considérer dans le développement d'une stratégie de préservation et d'archivage des revues savantes sur support numérique. Le problème peut paraître simple à première vue : il s'agit en somme de faire transiter des bits d'information vers le futur. Pourtant, le défi de la préservation de l'information numérisée ne sera relevé qu'avec l'adoption d'une stratégie impliquant plusieurs acteurs et portant autant sur les supports physiques que sur les formats d'encodage.

Obsolescence technologique

Les mesures de conservation visent à contrer le problème de l'obsolescence technologique, en premier lieu l'obsolescence physique. Pour comprendre le problème, on n'a qu'à évoquer les disquettes de 8 pouces, introduites sur le modèle 3310 d'IBM en 1979, ou même les disquettes de 5 pouces, introduites en 1980 sur les premiers ordinateurs personnels, qui sont illisibles avec les ordinateurs d'aujourd'hui. 20 ans à peine! on ne parle pas ici de textes écrits par des Sumériens!

Le rafraîchissement routinier des données assure de conserver le flot de bits en bon état, mais il faut également lutter contre l'obsolescence des logiciels : pensons ici au logiciel WordStar, fort populaire il

y a à peine 15 ans, pensons aussi aux multiples versions d'un même logiciel qui paraissent à intervalle régulier.

Les formats d'archivage et de conservation

Depuis les cinq dernières années, plusieurs projets cherchent à définir les meilleures pratiques pour l'archivage et la conservation des publications électroniques. Nous avons porté une attention particulière aux projets impliquant des publications scientifiques, à la question des formats et des supports, ainsi qu'aux stratégies employées pour l'archivage et la conservation.

Puisque les documents numériques sont stockés selon certains formats, il est nécessaire de s'interroger sur les critères de choix d'un format d'encodage permettant la représentation de l'information et la conservation à long terme.

Pour assurer la préservation de l'information, le format choisi doit être « lisible » par une application, et ce, aussi longtemps qu'il est souhaité. C'est ici que les difficultés se présentent.

Tout d'abord, il est important de mentionner que les conversions d'un format à l'autre ou encore d'une version à l'autre d'un même format ne sont pas une solution à ce problème. Prenons, par exemple, un format de traitement de texte comme Word. Un document stocké dans la version « 97 » de Word pourra être conservé pendant un certain temps, et nous pouvons supposer qu'il existera des applications qui pourront lire des documents Word pour encore un grand nombre d'années. Toutefois, ces logiciels auront évolué et même si en apparence le logiciel lira notre document, en fait il effectuera une conversion dans son format « natif ». Rien ne garantit que cette conversion fonctionnera correctement à tout coup. Les pertes d'information ou changements dans la présentation sont des situations courantes lors de telles conversions.

L'utilisation d'un format d'encodage de l'information à la fois simple et universel permet de pérenniser les documents numériques. Le SGML et, depuis 1998, le XML, par leur statut de norme et leur utilisation répandue, sont reconnus comme des formats stables d'encodage de l'information. Un des arguments en faveur du XML,

outre ses caractéristiques techniques et les possibilités d'exploitation intéressantes, est qu'il ne soit pas rattaché à un logiciel en particulier. Sa nature « non-propriétaire » en fait un format libre et ouvert, offrant une certaine garantie pour la préservation de l'information.

Le format XML peut être représenté à l'aide du jeu de caractères ASCII. Concrètement, le fichier produit sera un pur fichier ASCII, soit le type de fichiers le plus universel que l'on trouve dans le monde informatique. Il devrait exister, pour encore de nombreuses années, des plates-formes informatiques et des applications qui permettront de « voir » un fichier ASCII. Cette facilité de lecture par l'humain est impossible avec des formats binaires tels que Word, qui sont destinés à être compris que par une machine. Même si on perdait toute possibilité d'utiliser des applications pouvant faire un traitement intéressant de documents XML, un seul lecteur de documents ASCII permettra de consulter le document et de le comprendre.

Les documents structurés stockés en XML ont donc comme grande qualité d'être très bien adaptés pour la conservation à long terme, ce qui est fort intéressant dans le monde de l'édition scientifique. Des techniques, bien connues et maîtrisées, de rafraîchissement et de migration pourront être employées sans difficulté avec les documents en format structurés (XML, SGML), puisqu'ils ne contiennent que du texte « pur ». Pour assurer l'intégrité des documents qui contiennent des objets numériques (images, sons, modèles, formules, hyperliens, etc.), la même attention doit être portée à l'information « non-textuelle » qui constitue souvent une partie importante des articles de revues savantes. Ces différents objets liés peuvent être de formats propriétaires, incompatibles ou simplement de différentes versions.

Le tableau suivant présente les formats utilisés par huit programmes pour la préservation des documents électroniques.

Projet	Formats
Pandora (Preserving and Accessing Networked Documentary Resources of Australia; http://pandora.nla.gov.au/pandora/)	PDF, SGML, HTML
HighWire (États-Unis; http://highwire.stanford.edu/)	SGML, PDF
Muse (États-Unis; http://muse.jhu.edu/)	HTML, PDF
Allen Press (États-Unis; http://www.allenpress.com/)	SGML
Institute of Electrical and Electronic Engineers (États-Unis; http://www.ieee.org/)	SGML, PDF
American Astronomical Society (États-Unis; http://www.aas.org/)	SGML
American Institute of Physics (États-Unis; http://www.aip.org/)	PDF, SGML
Danemark, projet de dépôt légal des publications électroniques (http://www.pligtaflevering.dk/, http://www.konbib.nl/infolev/liber/articles/dupont11.htm)	ASCII Text, format d'image (par exemple TIFF)

Les formats les plus utilisés sont les HTML, SGML et PDF. Le format XML est encore peu utilisé pour le moment, mais il doit dorénavant être considéré. Le PDF est un format propriétaire, largement utilisé et accepté par le milieu de l'édition. Son accessibilité à long terme demeure toutefois source d'inquiétude pour les archives nationales et les bibliothèques (Hodge G. et Carroll B.C., 1999, 60). La garantie d'accès à long terme aux fichiers PDF est probable, mais ne peut être affirmée d'une façon aussi certaine.

On constate que les pratiques quant au choix du format d'encodage pour la préservation des fichiers textes vont clairement dans le sens de l'utilisation d'un format de balisage structuré normalisé, telle XML et le SGML.

Les supports d'archivage et de préservation

Les supports d'archivage et de préservation utilisés par les projets étudiés ne sont que très rapidement mentionnés, souvent ils ne le sont pas du tout. La pratique est de procéder à des copies de sécurité routinières sur divers supports magnétiques (ruban, cartouches, disques miroirs) ou optiques (cédérom, le DVD étant encore peu utilisé).

La question des supports mérite réflexion, car leur durée est comptée en dizaine d'années et non pas en siècle comme c'est le cas pour le papier de très haute qualité et les microfilms. Les supports magnétiques subissent une double détérioration. D'une part, l'affaiblissement progressif du champ magnétique nécessite un rafraîchissement périodique. D'autre part, les conditions environnementales (le taux d'humidité, les variations de température, la pollution, la poussière, etc.) contribuent à leur détérioration. Les supports optiques sont plus durables, mais ils restent sujets à la détérioration due à l'environnement (conditions ambiantes), aux matériaux utilisés pour leur fabrication, à la corrosion des différentes couches de métal, etc.

La stratégie pour assurer la préservation des documents numériques doit donc comprendre le rafraîchissement et la migration des données vers des supports diversifiés et fiables. De plus, comme pour le papier, la répartition géographique des supports constitue une condition importante pour la préservation des documents numériques.

Les stratégies employées dans les divers projets

Pour pallier la détérioration rapide des supports, les projets analysés utilisent différentes techniques, telles que la redondance des données, la dispersion géographique, les copies de sécurité de routine, sur une base quotidienne, de façon à ce qu'aucun événement, tels qu'une panne ou un bris du matériel, une attaque de hackers ou une catastrophe naturelle ne puissent détruire toutes les données.

Les bibliothèques ont aussi créé un modèle distribué d'archivage pour le matériel en ligne : LOCKSS (Lots of Copies Keeps Stuff Safe).

> *LOCKSS is a self-organizing, freeware-based, low-cost, voluntary approach to archiving online material, self-selected by participating institutions, that relies on consensus among several linked servers to determine authoritative states of files and restore lost or damaged files automatically (http://lockss.stanford.edu/projectdescbrief.htm).*

Cet outil permet aux bibliothèques de conserver les publications en ligne sur les disques des ordinateurs locaux. Les publications téléchargées localement ne sont pas effacées et sont continuellement confrontées aux mêmes publications qui sont en ligne pour s'assurer que leur contenu ne soit pas détérioré ou perdu, auquel cas, les publications sont restaurées. Plus les bibliothèques utiliseront ce modèle distribué assurant une répartition géographique des fichiers, plus il y aura de copies des publications conservées.

Il s'agit d'une approche intéressante, différente de celle des copies de sauvegarde, mais encore trop récente pour tirer des conclusions. Pour participer à un système comme LOCKSS, les producteurs de revues savantes doivent s'assurer de respecter les conditions techniques de compatibilité en évitant de définir des en-têtes qui limitent l'utilisation ou empêchent de déposer des documents dans la cache (Irreversible Publishing : Local control of content delivered via the Web. http://lockss.stanford.edu/projectdescfaq.htm).

La préservation comprend un site d'hébergement des documents offrant des assurances de sécurité physique et réseau à la fine pointe des technologies disponibles. Également, une redondance des données devra être assurée avec des partenaires choisis et répartis géographiquement. De plus, comme dans le cas du projet Muse, les abonnés institutionnels des revues devraient recevoir une copie sur cédérom de la version XML de la revue (sans le moteur de recherche) chaque année.

L'émulation comme stratégie de conservation à long terme

Différentes approches sont suggérées pour conserver les documents numériques à long terme : (1) l'application des normes et des standards pour les formats, (2) la conservation de la technologie (les logiciels et le

matériel informatique), (3) la migration des documents dans une forme accessible pour les générations futures, (4) l'impression des documents électroniques sur papier.

Aucune de ces solutions n'est totalement satisfaisante. Certains spécialistes préconisent une cinquième option : l'émulation.

L'émulation est une opération de simulation qui consiste à imiter le fonctionnement d'un ordinateur ou d'un logiciel sur un autre ordinateur généralement plus puissant d'une génération subséquente. On recrée virtuellement l'environnement matériel et logiciel d'origine. Les documents électroniques sont ainsi accessibles et lisibles sous leur forme originale.

Cette stratégie peut s'avérer intéressante, car elle n'altère pas les données. Elle permet de conserver l'aspect, le cachet et l'originalité du document numérique aussi bien que de son contenu. Aucune autre opération n'est requise si ce n'est le rafraîchissement du support sur lequel sont les données. Toutefois, cette solution demeure encore largement théorique et trop peu d'études ont été menées sur le sujet pour en tirer des conclusions. De plus, on peut s'attendre à ce que les coûts de recréation d'environnements technologiques complexes soient faramineux. Nous ne recommandons pas cette stratégie dans l'immédiat.

Garantir l'intégrité et l'authenticité des textes

L'authenticité et l'intégrité des textes sont deux éléments essentiels de la communication savante. Les auteurs autant que les lecteurs veulent s'assurer que les documents électroniques n'ont pas été manipulés, altérés ou encore falsifiés après leur création et leur publication.

Afin que les publications électroniques soient protégées et garanties, plusieurs solutions existent dont le cryptage. Le cryptage est, selon les termes de l'Office de la langue française, l'« opération par laquelle est substitué, à un texte en clair, un texte inintelligible, inexploitable pour quiconque ne possède pas la clé permettant de le ramener à sa forme initiale » (http://www.olf.gouv.qc.ca/). Essentiellement, il s'agit de coder un message de façon que seul un interlocuteur connaissant la « clé » puisse le décoder. Le cryptage apparaît comme une solution

adéquate pour la préservation de l'authenticité et de l'intégrité des publications électroniques scientifiques. Cependant, certaines difficultés pour la conservation des documents à long terme sont prévisibles, par exemple, dans le cas où un éditeur cesse ses activités et que la clé, par la même occasion, est perdue.

Signalons une autre solution, présentement en développement par un groupe de travail du World Wide Web Consortium, le Digital Signature Initiative (Dsign ; http://www.w3.org/Signature/), qui a comme mandat de développer une syntaxe XML représentant la signature des ressources Internet. Ces signatures permettront d'assurer l'intégrité des données, de même que l'authentification. À terme, l'objectif est d'établir le « Web of trust ».

Nous proposons d'opter pour la simplicité et l'élégance de cette dernière solution et chercher à implanter la signature numérique dès que la norme aura atteint le statut de recommandation du W3C.

La responsabilité de la conservation et de l'archivage

Généralement, par la loi du dépôt légal, les bibliothèques nationales s'assurent que toutes les publications de leur pays soient acquises, conservées et diffusées. Toutefois, dans plusieurs pays, la législation sur le dépôt légal ne couvre pas encore les publications diffusées en réseau. Comme il n'existe pas encore d'infrastructures nationales reconnues pour assurer l'archivage et la conservation à long terme de ces publications électroniques, qui doit assumer cette responsabilité ?

Cette question est étudiée dans plusieurs pays. Éventuellement, les lois seront modifiées pour inclure les publications en réseau. Dans l'attente d'une nouvelle législation, le dépôt des publications électroniques diffusées en réseau se fait sur une base volontaire dans de nombreux pays, notamment au Canada, en Australie, en France, en Suisse, au Royaume-Uni et en Allemagne (Libby M., 1999). Plusieurs rapports, dont celui du Task Force on Archiving of Digital Information (http://www.rlg.org/ArchTF/tfadi.index.htm), suggèrent que les éditeurs soient les premiers responsables de l'archivage et de la conservation de leurs publications. En raison de la complexité de ces tâches, et de leurs coûts inhérents, les éditeurs de revues savantes

auraient avantage à se doter d'une infrastructure commune qui aurait mandat d'assurer la conservation et l'archivage en relation avec les bibliothèques nationales lorsque celles-ci auront implanté le dépôt légal des documents en réseau.

Certains portails assument actuellement des responsabilités d'archivage, mentionnons JSTOR, dont les frais d'abonnements couvrent ce service. Pour les éditeurs de revues savantes, le modèle de JSTOR est intéressant et devrait faire partie d'une stratégie en collaboration avec les bibliothèques nationales et les éditeurs de revues.

La préservation implique une veille technologique constante afin d'assurer la mise à jour de la stratégie pour l'archivage de la version électronique des revues savantes. Là encore, comme pour les documents papier, une expertise et des pratiques professionnelles se développent pour garantir l'accès au patrimoine des revues savantes par les générations futures.

Glossaire des termes techniques

DHTML Le Dynamic HTML est la combinaison du HTML, de feuilles de styles et de scripts qui permet de créer des documents à la volée selon des besoins ou des caractéristiques précises, éventuellement identifiés par le lecteur.

DTD La définition de type de document est la « grammaire » d'un genre de texte (article, livre, dictionnaire, etc.) dans laquelle on retrouve la description des éléments, de leurs contenus et des relations entre les éléments.

GIF Format (Graphics Interchange Format), créé par CompuServe en 1987 et amélioré en 1989, utilisé en général pour les figures ; très utilisé dans le Web, il ne permet d'utiliser que 256 couleurs ou tons de gris, mais génère des fichiers très légers.

HTML Format (HyperText Markup Language) est une application simple du SGML. Facile à apprendre, ce format de diffusion est le plus utilisé dans le Web. Ce format présente un jeu de balises limité.

JPEG Format (Joint Photographic Expert Group) utilisé en général pour les figures qui permet l'utilisation de 16 millions de couleurs ; surtout employé pour représenter des photographies, ce format est reconnu pour son excellent algorithme de compression.

RDF Resource Description Framework (RDF) est une création pour le traitement des métadonnées ; il fournit l'interopérabilité entre les applications qui échangent de l'information non compréhensible par les machines sur le Web. RDF augmente la facilité de traitement automatique des ressources Web.

ROC (OCR) Reconnaissance optique de caractères (Optical Character Recognition). Les logiciels de ROC permettent de reconnaître les caractères à partir d'une image d'un texte et d'en produire une version textuelle.

PDF Format propriétaire introduit par Adobe (Portable Document Format) qui peut être lu sur plusieurs plateformes ; il préserve l'apparence originale du document et permet l'intégration d'hyperliens, de signets et de métadonnées. Format particulièrement intéressant pour l'impression à distance.

PNG Le PNG (Portable Network Graphics) est la nouvelle étoile filante des formats d'images. Il est destiné à remplacer le GIF et le TIFF. Il s'agit également d'un format non propriétaire.

Schémas XML Modèles de documents qui permettent d'exprimer l'ensemble des contraintes que doit respecter un document XML. Ils permettent la validation des contenus. Destinés à remplacer la DTD dans un système XML.

SGML Langage structuré (Standard Generalized Markup Language) normalisé par l'ISO en 1986 ; il s'agit d'un métalangage qui permet de décrire la structure logique d'un document.

TIFF Format de type image en mode point (Tagged Image File Format) développé par Aldus et Microsoft ; reconnu par les archivistes et plusieurs guides des meilleures pratiques pour l'archivage des documents en formats images en raison de son algorithme de compression qui assure aucune perte d'information.

XHTML Reformulation (eXtensible Hypertext Markup Language) du HTML 4.0 en XML : une passerelle pour favoriser le passage du HTML au XML.

Xlink Encore à l'étude par le W3C, XLink (XML Linking Language) est la norme qui décrit la façon d'intégrer des liens hypertextes à un fichier XML. Le XLink permet notamment de qualifier la nature des liens.

XML Langage de balisage structuré, basé sur le SGML, XML (eXtended Markup Language) a depuis 1998 le statut de recommandation du W3C (World Wide Web Consortium).

XSL Langage (eXtensible Stylesheet Language) permet de développer des feuilles de styles pour la représentation à l'écran des documents.

XSLTL XSLTLangage de transformation (eXtensible Stylesheet Language Transformation) qui permet de faire la conversion d'un document XML à un autre type de document XML.

Références

Boismenu G. et Beaudry G. (1999), « Publications électroniques et revues savantes : acteurs, rôles et réseaux. » *Documentaliste. Sciences de l'information*, Vol. 36, nº 6, décembre, p. 292-305.

Capian P. et W.Y. Arms (1999), « Reference Linking for Journal Articles », *D-Lib Magazine*, july/august, [consulté en décembre 2000]

http://www.dlib.org/dlib/july99/capian/07caplan.html

Chahuneau F. (2000), « XML : un langage universel pour la représentation textuelle de données structurées », *Bibliothèques numériques, Institut national de recherche en informatique et en automatique*, ADBS Éditions, p. 119-141.

Chartron G. (2000), « L'édition scientifique face à Internet », *Bibliothèques numériques, Institut national de recherche en informatique et en automatique*, ADBS Éditions, p. 189-227.

Clément C. et Bonvin M. (2000), *Les périodiques électroniques en sciences humaines et sociales*, Travail fait sous la direction de Françoise Khenoune. Bibliothèque cantonale et universitaire de Lausanne-Dorigny, Mars, http://www.unil.ch/BCU/recherch/l_art_bi.htm

Cleveland G. (1999), *Selecting Electronic Document Formats*, Juillet. http://www.ifla.org/VI/5/op/udtop11/udtop11.htm

Creating Digital Ressources for the Visual Arts : Standards and Good Pratice, consulté le 22 décembre 2000, http://vads.ahds.ac.uk/guides/creating_guide/sect32.html

Dahn M. (2000), « Counting Angels on a Pinhead : Critically lnterpreting Web Size Estimates », Online, January/February, p. 35-40.

Early Canadiana Online/Notre mémoire en ligne, consulté le 22 décembre 2000, http://www.canadiana.org

Febvre L. et Henri-Jean Martin, *L'apparition du livre*, Paris, Michel, 1971.

Georgia Institute of Technology (1998), Graphic, Visualization and Usability Center, *Tenth User Survey*, http://www.gvu.gatech.edu/user_surveys/survey-1998-10/

Grabmeier J. 2000, « Texts on Computer Screens Harder to Understand, Less Persuasive », Research News, août, http://www.acs.ohio-state.edu/units/research/archive/comptext.htm

Gray M., MIT, Internet Statistics ; *Growth and Usage of the Web and the Internet*, consulté le 22 décembre 2000, http://www.mit.edu/people/mkgray/net/

Haigh S. (1999), « Le projet de métadonnées Dublin Core », *Flash Réseau*, nº 63, décembre. http://www.nlc-bnc.ca/pubs/netnotes/fnotes63.htm.

Hodge G. et Carroll B.C. (1999), *Digital Electronic Archiving : The State of the Art and the State of the Practice*. International Council for Scientific and Technical Information, Information Policy Committee, 26 avril.

Hudgins J. et al. (1999), *Getting mileage out of Metadata, Application for the Library*, Chicago, American Library Association.

Irreversible Publishing: Local control of content delivered via the Web. consulté le 22 décembre 2000, http://lockss.stanford.edu/projectdescfaq.htm

Lawrence S.C. et Lee G. (1999), « Accessibility of Information on the Web », *Nature*, vol. 400, 8 juillet. Http://www.nature.com/

Lecomte D. et al., *Les normes et les standards du multimédia*, Paris, Dunod, 1999.

À Basic Introduction to PNG Features, consulté le 22 décembre 2000, http://www.libpng.org/pub/png/pngintro.html

Libby M. (1999), *Gestion des publications électroniques diffusées en réseau : état de la question dans divers pays*, Bibliothèque du Canada, 31 décembre.

Liberatore K. (2000), « Getting to the Source », *Macworld*, consulté le 22 septembre. http://macworld.zetnet.com/features/pov.4.4.html/

Lupovici C. (1998), « L'information bibliographique des documents électroniques », *Bulletin des Bibliothèques de France*, t-43, nº 4, 1998, p.42-48

Lynch P.J. et Horton S. (1997), *Yale Style Manual – Site Design*, consulté le 22 décembre 2000, http://info.med.yale.edu/caim/manual/contents.html

Morgan C. (1999), « Journals metadata : information about content », *Learned Publishing*, Vol. 12, nº 3 julliet, p. 191-195.

Murphy K. (1996), « Cheaters Never Win« , Internetworld, Mai. http://www.internetworld.com/print/1996105/20/undercon/cheaters.html/

Netcraft (2000), *November 2000 – Web Server Survey*. consulté le 22 décembre 2000, http://www.netcraft.com/survey/Reports/0011/

Nikitenko C. (2001), « De l'édition traditionnelle à l'édition numérique », *Les cahiers du numérique*, vol. 1, nº 5, p. 19-44.

Ockerbloom J.M. (2001), « Archiving and Preserving PDF Files », *RLG DigiNews*, February. http://www.rlg.org/preserv/diginews/

Salaun J.-M. (2000), « Du partage des collections à la fourniture de documents : nouvelles relations entre éditeurs et bibliothécaires », *Bibliothèques numériques, Institut national de recherche en informatique et en automatique*, ADBS Éditions, p. 99-118.

Teasdale G. (2001), 'Vers l'e-édition savante ? », *Les cahiers du numérique*, vol. 1, nº 5. p. 167-182.

Tracey S (1997), « Moving Up The Rank« , *Ariadne*, nº 12, novembre. http://www.ariadne.ac.uk/issues12/search-engine/.

Tracey S. (1997), « Keyword Spamming : Cheat Your Way To The Top », *Ariadne*, nº 10 juillet. http://www.ariadne.ac.uk/issues10/search/engines/

Tracey S. (1997), « The relevance of underpants to seaching the Web », *Ariadne*, nº 24, juillet. http://www.ariadne.ac.uk/issues24/search/engines/intro.html

Van de Sompel H. (1999), « Reference Linking in a Hybrid Library Environment », *D-Lib Magazine*, April, http://www.dlib.org/dlib/a99/caplan/07caplan.html

Vézina, M.-H., et M. Sévigny (2000). « De l'imprimé vers l'électronique : réflexions et solutions techniques pour une édition savante en transition », *Documentaliste. Sciences de l'information*, Vol. 36, nº 6, décembre, p. 306-320.