LA GESTION QUOTIDIENNE DES PÉRIODIQUES ÉLECTRONIQUES

Marc-Henri BAWIN
Ulg - Cellule Internet
Place Cockerill, 1 - Bât. A3 - B-4000 Liège
email : MH.Bawin@ulg.ac.be
Philippe MOTTET
Ulg - U.D. Walthère Spring
Institut de Chimie B.6 - B-4000 Sart Tilman (Liège 1)
email : pmottet@ulg.ac.be

I. ACCÈS AUX REVUES

L'apparition d'Internet et sa facilité à diffuser l'information de façon électronique ont sensiblement modifié le monde de l'édition et particulièrement le domaine des journaux scientifiques. Jusqu'au milieu des années 1990, ces revues n'ont pratiquement connu qu'un seul support : le format imprimé. Il est vrai que les éditeurs proposent depuis longtemps déjà certaines informations sur des supports électroniques comme les disquettes ou les CD-ROMs, mais cette pratique reste marginale et limitée à des fins bien précises :

- les disquettes servent habituellement à contenir des informations complémentaires aux articles diffusés dans les journaux imprimés, comme des listes de données ou des résultats d'analyses, et semblent par leur faible capacité de volume condamnées à moyen terme;
- les CD-ROMs ont comme principal avantage, y compris pour l'avenir, d'être un substitut valable au format imprimé pour l'archivage des années écoulées. C'est déjà l'option choisie par les éditeurs qui proposent leurs journaux sous cette forme puisque les CD-ROMs recouvrent souvent plusieurs fascicules de la version imprimée et ne sont généralement publiés qu'une, deux ou quatre fois par an.

Sur Internet par contre, la mise sous forme électronique des journaux scientifiques a connu une véritable explosion depuis deux ou trois ans. Le nombre de journaux scientifiques imprimés a augmenté de façon exponentielle en quelques décennies : ils étaient environ 100 au début du 19e siècle, 10.000 en 1950 et plus de 70.000 en 1987.

En 1995, moins de 200 journaux possédaient un site sur Internet. Aujourd'hui on les y trouve pratiquement tous, du moins dans les domaines où la littérature périodique est prédominante comme les diverses branches de la médecine, des sciences ou des sciences de l'éducation, avec les conséquences que l'on imagine sur la gestion des bibliothèques mais aussi bientôt sans doute sur le comportement des scientifiques

1. Avantages et inconvénients des journaux électroniques

La possibilité d'accéder en ligne aux revues scientifiques comporte beaucoup d'avantages pour le chercheur. Dans le meilleur des cas, il peut :

- les consulter directement de son bureau, voire de chez lui, en même temps que des milliers d'autres lecteurs et au moment même où elles sont publiées puisqu'il aura reçu quelques heures auparavant par courrier électronique l'annonce de la mise en accès du dernier numéro;
- effectuer des recherches par mots-clés ou par auteur;
- sauter d'une référence à une autre grâce aux liens hypertextes;

- afficher séparément ou agrandir des zones de texte, des graphiques, des tableaux;
- profiter des possibilités qu'offrent les logiciels de sons ou d'animation d'images;
- et cela à toutes les heures du jour et de la nuit puisque les horaires d'ouverture de sa bibliothèque ne sont plus une contrainte.

Le bibliothécaire aussi peut y trouver son compte, ne fût-ce que dans la résolution des problèmes d'espace, de dégradation ou de vol et dans celle des problèmes causés par les fascicules non reçus.

Pourtant la situation n'est pas aussi simple qu'on pourrait le croire, tout au moins dans la période de transition que nous connaissons, période où s'implante un nouveau media mais où les bibliothèques n'ont pas renoncé à leurs abonnements imprimés et rechignent, à juste titre, à investir dans ces suppléments électroniques.

Tout d'abord, les avantages que nous venons d'énumérer ne faciliteront l'accès à l'information scientifique que dans des conditions d'utilisation optimales, et tous les familiers des ordinateurs connaissent les désagréments que l'informatique peut causer : les connexions sont quelquefois très lentes ou même temporairement indisponibles, les liens hypertextes peuvent renvoyer à des pages désormais périmées, posant ainsi le problème de la pérennité des adresses URL et de la consultation des documents des années écoulées. Par ailleurs, le matériel informatique et ses accessoires ne sont pas non plus à l'abri des vols et des dégradations.

D'autres questions restent pour l'instant sans réponse. Elles concernent bien sûr l'aspect financier de cette évolution ou la préservation de l'authenticité de documents désormais aisément falsifiables, mais aussi certaines notions traditionnelles de la littérature périodique, comme celles de volume ou de fascicule qui pourraient bien n'avoir plus aucun sens à l'avenir : certains journaux déjà diffusent sur leur site des articles au fur et à mesure de leur acceptation sans tenir compte d'aucun suivi dans la numérotation.

Pour l'utilisateur, les différents avantages que nous venons de citer sont souvent contrebalancés par des inconvénients d'ordre pratique, liés au manque d'habitude à utiliser ce nouveau media :

- il faut d'abord, bien sûr, disposer d'un ordinateur connecté à Internet, ce qui est loin d'être encore généralisé. On ne fait que déplacer le problème si l'on estime que c'est dans les bibliothèques que doivent se trouver les terminaux nécessaires à la consultation des journaux en ligne, ce qui amènerait une surcharge financière importante et récurrente tout en annulant les avantages de la consultation directe et permanente;
- d'autre part, le chercheur ignore en général les titres qui sont disponibles et ce qu'on trouve sur les sites des journaux;
- il doit aussi charger sur son disque dur les logiciels (viewers, plug-ins) nécessaires à la lecture et à l'impression des articles et il doit en apprendre le fonctionnement;
- enfin, il doit surmonter une méfiance assez naturelle à soumettre des papiers à des revues uniquement électroniques qui n'ont pas encore obtenu la reconnaissance scientifique des journaux traditionnels.

Comme on le voit, le chercheur est en fait très dépendant du bibliothécaire qui voit son rôle évoluer et devenir celui d'un intermédiaire de plus en plus actif entre le fournisseur et le consommateur d'informations.

2. Informations disponibles sur le Web

Dans la gestion quotidienne de ces périodiques, la situation n'est pas non plus toujours simple : les journaux électroniques ne fournissent pas tous les mêmes informations et ne proposent pas tous un accès identique. Quels types d'information

peut-on trouver sur Internet et comment trouver les sites des journaux ?

2.1. Types d'information

C'est l'aspect le plus ambigu des journaux électroniques. Dans l'état actuel des choses, il semble que beaucoup d'éditeurs considèrent encore Internet comme une simple vitrine permettant surtout de promouvoir leurs journaux imprimés ou bien ne possèdent pas encore parfaitement les possibilités techniques offertes par ce nouveau moyen de communication. Quoi qu'il en soit, et même si l'on peut supposer que la situation va peu à peu se stabiliser, il faut reconnaître que les informations proposées actuellement sur les différents sites varient assez fortement d'un journal à l'autre. On peut les ranger en 4 catégories :

- les informations générales (domaine couvert par la revue, membres de l'editorial board, prix et modalités d'abonnements, instructions aux auteurs,...);
- les tables des matières, ou ToC (Tables of Contents);
- les résumés des articles (abstracts);
- le texte complet (full-text) de l'ensemble des articles de la revue, d'articles sélec-tionnés ou du dernier fascicule paru.

Dès lors qu'un journal se trouve sur Internet, il offre gratuitement au minimum le premier type d'information, ce qui semble logique puisqu'il s'agit là d'indications souvent nécessaires à la promotion de la revue.

L'accès aux autres informations, ToC, abstract et full-text, varie selon les éditeurs et parfois selon les titres de journaux d'un même éditeur. Les Tables of Contents sont désormais presque toujours en accès libre mais les années proposées sont très variables, parfois seulement l'année en cours, dans d'autres cas une dizaine d'années ou plus.

En ce qui concerne les abstracts et les full-texts, si l'on imagine comme finalité à moyen terme pour chaque journal un accès payant et peut-être même le remplacement pur et simple de la version imprimée, on peut dire qu'on se trouve dans une période intermédiaire où toutes les combinaisons sont actuellement possibles : ToC mais pas abstracts, ToC et abstracts mais pas full-text, full-text limité à certains articles où à certaines périodes, gratuité totale, gratuité limitée à une période d'essai, gratuité pour les souscripteurs à la version imprimée, ou accès payants qui eux-mêmes peuvent se présenter de diverses manières. En général, on s'abonne à la version imprimée avec la possibilité, moyennant supplément, d'obtenir un accès à la version en ligne. Mais il est également possible de ne souscrire qu'à la version en ligne et il existe aussi l'obligation de prendre un abonnement à la fois à la version imprimée et à la version en ligne. Cela ne prête pas à conséquence quand cet abonnement est ajouté gratuitement ou sans augmentation notable à l'abonnement imprimé, mais cela ressemble furieusement à de la vente forcée quand le prix de l'abonnement est augmenté en conséquence.

Notons enfin qu'à côté du format électronique de journaux publiés parallèlement à leur forme imprimée, de plus en plus de nouvelles revues en version uniquement online apparaissent avec les mêmes caractéristiques que leurs homologues imprimés (spécialisées dans un domaine bien particulier, patronnées par des scientifiques de haut niveau,...) et avec la même ambition de recevoir la reconnaissance de valeur et de qualité par la communauté scientifique (articles contrôlés par des pairs, facteurs d'impact,...).

On le voit, la gestion sur Internet d'un nombre plus ou moins élevé de journaux scientifiques devient une tâche en soi, et l'on comprend que les chercheurs évitent de s'aventurer dans ce dédale de possibilités et laissent ce soin à leurs bibliothécaires.

2.2. Accès aux sites

En général, on accède à une revue sur Internet via le site de l'éditeur (*ACQWEB*

répertorie un grand nombre d'adresses électroniques d'éditeurs sur le site **http://www.library.vanderbilt.edu/law/acqs/pubr.htlm**
Mais il n'est pas toujours aisé de connaître l'éditeur de la revue que l'on recherche et c'est l'une des raisons qui a poussé un grand nombre de bibliothèques à créer leur propre page Web répertoriant les titres des journaux et leur adresse URL dans le domaine qui les concerne. Sans aller très loin, on peut signaler que les universités belges ont toutes sacrifié à cette pratique, ne se différenciant entre elles que par le nombre d'entrées proposées ou par la présentation choisie : ordre alphabétique des titres des journaux, classement par disciplines ou par types d'infor-mations disponibles. Il est en outre sou-vent indispensable de passer par la page de sa bibliothèque pour accéder aux re-vues autorisées dans son institution.

Si la revue ne se trouve pas dans une de ces listes, il existe des répertoires généraux ou spécialisés sur Internet :

- *ARL (Association of Research Libraries)* pour la 6e édition (une 7e est en cours de réalisation) à l'adresse : **http://arl.cni.org/scomm/edir/6th/a.html**
- *Library Online SWT* à l'adresse : **http://www.library.swt.edu/ejs/**
- *NewJour* à l'adresse : **http://gort.ucsd.edu/newjour**
- *Alliance* à l'adresse : **http://www.coalliance.org/ejournal/**
- *E-DOC* à l'adresse : **http://www.edoc.com/ejournal/**
- *Korea Basic Science Institute* à l'adresse : **http://biblio.kbsi.re.kr/yellow/index.html**
- *CLIC Consortium Electronic Journal Project* (pour la chimie) à l'adresse : **http://www.ch.cam.ac.uk/CLIC/clic-lib.html**

En dernier recours, les moteurs de recherche du type *AltaVista*, *Lycos* ou *Yahoo* donnent presque toujours satisfaction.

Signalons enfin que certaines sociétés se sont fait une spécialité de la gestion globale des périodiques d'une bibliothèque ou d'une institution et proposent des contrats incluant l'accès en ligne vers les journaux autorisés :

- *Ebsco Online* à l'adresse : **http://www.ebsco.com/home/bropage1.asp**
- *InformationQuest* à l'adresse : **http://www.eiq.com/**
- *SwetsNet* à l'adresse : **http://www.swetsnet.nl/**

3. Modalités d'accès

Gratuit ou non, l'accès à un journal électronique se présente généralement de la manière suivante : la page d'accueil de la revue conduit au moyen de liens hypertextes vers les tables des matières des années consultables, et de là vers le résumé et le texte des articles. Le format de celui-ci est précisé et le chargement éventuel sur son disque dur du logiciel nécessaire à sa lecture est souvent possible à partir du site même du journal.
Quand il n'est pas gratuit, l'accès à un site Internet est limité aux utilisateurs autorisés de différentes façons : mot de passe, numéro d'identification, formulaire d'enregistrement avec autorisation de ponctionner sa carte de crédit,...
Dans le domaine des journaux scientifiques, les éditeurs imposent des contrats et des licences qui utilisent principalement 2 systèmes de limitation d'accès à leur site.

3.1. Restriction par adresses IP

La façon la plus simple et la plus sûre pour limiter un accès est de le restreindre à certaines machines, en fonction de leur adresse électronique.

Chaque ordinateur connecté à Internet possède une adresse IP qui est unique et qui se présente sous la forme X.X.X.X. où X est compris entre 0 et 255. Une adresse de classe B identifie les machines dont les deux premiers segments sont communs. Les trois premiers constituent la classe C de l'adresse et le quatrième segment identifie un ordinateur particulier.

L'attribution des adresses se fait notamment en fonction de la taille des institutions et des sociétés : dans le cas d'une entreprise n'utilisant que quelques dizaines de machines au maximum, une adresse de classe C sera suffisante puisque les trois premiers segments pourront être communs à toutes les machines. Par contre, quand une institution regroupe plusieurs centaines d'ordinateurs connectés à Internet, il est alors indispensable de lui attribuer plusieurs classes C et seuls les deux premiers segments identifieront de façon univoque les machines de cette institution. Si l'on prend l'exemple d'un ordinateur de l'Université de Liège connecté à Internet et possédant l'adresse 139.165.204.n, il a en commun avec toutes les autres machines de l'Université les deux premiers segments de son adresse, soit 139.165. Celles qui ont 204 comme troisième segment sont beaucoup moins nombreuses (256 au maximum, quelques-unes en réalité). Ce sont celles qui font partie d'un même noeud, c'est-à-dire, d'un même sous-réseau au sein de l'institution. Rien que pour l'Institut de Chimie de l'Université de Liège par exemple, il existe 5 ou 6 nœuds différents.

Lorsque le droit d'accéder à leur journal a été confirmé, certains éditeurs ouvrent l'accès à leur site pour toutes les machines ayant une adresse B commune. C'est la situation la plus confortable puisque l'ensemble de l'institution peut accéder librement au site en question, avec parfois la nécessité d'introduire un mot de passe, mais qui peut être connu de tous puisque la restriction se fait au niveau de l'adresse. La situation la plus complexe est celle où les éditeurs limitent leur accès aux adresses de classe C : quelques machines seulement auront alors la possibilité de se connecter à la revue. Ces restrictions sont généralement dues à la complexité des termes des licences accordées par les éditeurs et à leur méconnaissance des systèmes de réseaux qui peuvent varier largement d'une institution à l'autre. La seule solution dans ce cas est la négociation, parfois rude, avec l'éditeur pour tenter d'élargir l'autorisation à un plus grand nombre d'ordinateurs.

3.2. Restriction par mots de passe

D'autres éditeurs attribuent aux utilisateurs autorisés un numéro d'identification et/ou un mot de passe qu'il faut introduire à chaque consultation. Ce système est à la fois plus contraignant et moins fiable, du moins dans le cas d'un abonnement institutionnel : le bibliothécaire contrôle le droit d'accès des utilisateurs puisque c'est lui qui possède le numéro et le mot de passe, et cela rend obligatoire le passage par ce même bibliothécaire pour le chercheur qui souhaite accéder à la revue. Rien n'empêche ensuite l'utilisateur de divulguer les informations à qui il veut.

Signalons aussi qu'à côté de la consultation directe des journaux via les éditeurs, on trouve de plus en plus de systèmes sur Internet qui permettent d'effectuer des recherches bibliographiques et de commander automatiquement (de façon électronique ou par courrier traditionnel) les articles sélectionnés, soit par le biais de sa bibliothèque, soit par payement individualisé (carte bancaire ou compte déposé). Tous ces systèmes sont bien entendu payants et souvent très coûteux.

Ces systèmes peuvent être spécialisées comme :

pour *SciFinder* à l'adresse :
http://www.cas.org/SCIFINDER/ (domaine de la chimie)
ou *Medline* à l'adresse :
http://www4.ncbi.nlm.nih.gov/PubMed/
(domaine biomédical),
ou plus généraux comme les systèmes :
Inside de la British Library à l'adresse :
http://www.bl.uk/online/inside/
UnCover à l'adresse :
http://uncweb.carl.org/
ou ISI Web of Science à l'adresse :
http://www.isinet.com/products/citation/
wos.html.

II. FORMATS DE LECTURE

L'affichage des articles des journaux électroniques peut se faire dans différents

formats, reconnus directement par le navigateur ou nécessitant l'utilisation de logiciels spécialisés. Ces logiciels sont en général imposés par le fournisseur, c'est-à-dire l'éditeur, mais il peut arriver que le lecteur ait le choix entre plusieurs types de formats de lecture différents.

Il est difficile d'anticiper les types de format de lecture qui s'imposeront pour les journaux électroniques. Les pages suivantes proposent un tour d'horizon de l'ensemble des possibilités actuelles. Certaines d'entre elles ne sont pas, ou pas encore, utilisées dans le domaine qui nous occupe, d'autres se présentent déjà comme des standards pour l'avenir.

Documents prêts pour une publication en format électronique quelconque vs documents déjà publiés sous forme papier

1. Documents déjà publiés sous forme papier

Dans ce premier cas, la production du document est censée partir d'un document déjà publié sous forme papier. Il s'agit uniquement du transfert d'un support vers un autre, sans valeur ajoutée à l'information présente. Ce passage à la version électronique comprend deux étapes : la numérisation proprement dite implique l'utilisation d'un scanner (à main ou plus vraisemblablement à plat, à tambour pour les gros volumes de données); le décryptage du texte intelligible et codable pour un ordinateur à partir de l'image obtenue vient ensuite.

1.1 Documents images vs documents textes

On appelle documents images les fichiers obtenus juste après scannage (documents bruts), ou après un très léger traitement (documents comprimés), et qui contiennent le document sous forme codée : des points images (pixels) dont on précise les caractéristiques de luminance (plus ou moins clair) et de chrominance (d'un mélange de certaines couleurs). Les fichiers textes réclament un traitement supplémentaire.

1.1.1. Documents images

A. Documents bruts

- format TIFF (Tagged Image File Format)

Le format de fichier TIFF est le format traditionnel des images lorsqu'elles sortent du processus de scannage. Il accepte 16,7 millions de couleurs et n'est pas comprimé. De par ses caractéristiques, il est bien adapté à l'échange de documents entre intervenants - par exemple entre un opérateur de scanner et un infographiste qui va, à l'aide de logiciels adaptés, retoucher l'image pour en modifier les caractéristiques externes (plages de couleurs, sujets montrés, recadrage voire photo-trucage pur et simple) ou internes (définition, taille du fichier, nombre de couleurs, ...) - mais guère à l'envoi par l'opérateur de scanner vers l'utilisateur final de ce fichier, vu sa taille en octets. Pour ce genre d'échange, il est bon de réduire la taille du fichier, et donc, sur un réseau, le temps de téléchargement, en compressant le document. On utilise pour cela d'autres formats de fichiers, ils sont dits formats avec compression.

Les fichiers TIFF peuvent être lus avec n'importe quel logiciel de traitement d'images en mode point (donc non vectorielles) comme Photoshop, Paintshop Pro ou encore LView Pro.

B. Documents comprimés

a) format GIF (Graphics Interchange Format)

Le format de fichier GIF diminue sensiblement la taille du document (de 10 à 30%). Il est utilisé pour la compression de dessins au trait affichés sur des pages Web. La toute grosse majorité des icônes que l'on voit apparaître sur les sites Web graphiques adoptent ce format. Ses caractéristiques sont les suivantes :

- les fichiers acceptent un nombre maximal de 256 couleurs choisies parmi 16,7 millions;
- l'algorithme de compression est dit sans perte. Toute l'information binaire présente dans le fichier à l'entrée peut être intégralement retrouvée dans le fichier à la sortie. En appliquant l'algorithme de compression " *à l'envers* ", on peut retrouver l'image telle qu'elle était avant traitement;
- le format de fichier GIF est la propriété de CompuServe Inc., le fournisseur d'accès et prestataire de services américain, mais l'algorithme de compression des données utilisé pour GIF est la propriété de la société Unisys. Cet algorithme (LZW pour Lempel Ziv Welch) est basé sur deux types d'algorithme de compression de données sans perte : LZ77 et LZ78, mis au point respectivement en 1977 et 1978 par deux chercheurs israéliens, Jacob ZIV et Abraham LEMPEL. Terry WELCH, travaillant à l'époque pour la société Sperry (plus tard Unisys), en a perfectionné en 1983 une version beaucoup plus rapide. Deux chercheurs d'IBM, Victor MILLER et Mark WEGMAN, en firent autant presque au même moment. Tous trois obtinrent un brevet pour cet algorithme. Or celui-ci est utilisé dans de nombreuses techniques de compression/décompression de données comme ici GIF, mais aussi TIFF-LZW (une version compressée de TIFF), PostScript (langage de description de pages dont nous parlerons plus loin), Portable Document Format (PDF, format de fichier mis au point chez ADOBE), V.42bis (norme de compression de données utilisée dans les modems pour parvenir à une vitesse de 33600 bits par seconde), et d'autres.. Il s'agit donc d'un format " *propriétaire* " pour lequel CompuServe Inc. et Unisys peuvent demander des royalties. Actuellement tous les producteurs de logiciels, y compris Microsoft, implémentant d'une manière quelconque cet algorithme doivent signer un accord de licence avec Unisys. Cette situation est gênante et, n'était le fait que ce format est déjà très largement utilisé sur Internet, il devrait être abandonné en raison des risques de non pérennité, au profit d'un format " *libre* " ou, du moins, utilisant un algorithme de compression dont les sources sont libres ; comme ceux qui suivent;
- une possibilité intéressante de ces fichiers est ce que l'on appelle l'entrelacement. Il consiste à ne pas afficher toute l'image d'un coup mais à le faire pour certaines lignes seulement. Une première passe dessine environ 12,5 % de l'image sur la totalité de sa surface donnant une sorte d'image en très basse définition mais permettant déjà d'en distinguer vaguement le sujet. L'image est alors redessinée en trois passes successives en ajoutant, respectivement, 12,5%, 25% et 50 % de l'information contenue dans l'image;
- il faut encore distinguer deux types de fichiers GIF: le GIF87a et le GIF89a. Ce dernier permet d'inclure dans un seul fichier une succession d'images. Si ces images sont bien choisies et disposées, elles s'enchaînent pour donner l'illusion d'un petit film : ce sont les " *gifs animés* ". Il permet aussi de définir une couleur, et une seule, comme couleur " *transparente* ". Celle-ci est définie à la sauvegarde et à l'affichage, les zones de l'image qui ont cette couleur laisseront apparaître la teinte qui se trouve au-dessous.

Les fichiers GIF peuvent également être visualisés dans un logiciel de traitement d'images mais sont, plus simplement, affichables par n'importe quel browser Internet (sauf bien sûr ceux orientés vers le texte exclusivement).

b) format JPEG (Joint Photographic Experts Group)

Le format de fichier JPEG permet de diminuer considérablement la taille des documents. Il est utilisé pour la compression des photos et permet de " *choisir* " le taux de compression. Voici ses caractéristiques :

- les fichiers possèdent obligatoirement 16,7 millions de couleurs, soit l'utilisation de 24 bits pour coder la couleur

de chaque pixel, ce qui semble énorme (3 fois ce qui est utilisé pour les images GIF) et *a priori* pénalise la compression. Il n'en est rien, car l'algorithme de compression utilisé est tout autre;

- l'algorithme de compression est dit " *avec perte* ", car en reprenant une image comprimée, il sera impossible de reconstituer l'image de départ. Cette perte est due au fait que l'algorithme peut coder en une fois des lignes horizontales de points ayant la même teinte, ce qui prend beaucoup moins de place. Plus les chaînes de points de couleur seront longues et plus le fichier gagnera en taille, mais plus ces séries seront longues et plus l'image comprimée sera différente de l'image de départ.

  Ce paramètre peut être fixé par l'utilisateur qui effectue un choix entre qualité d'image et taille du fichier. Pour de faibles taux de compression, les différences entre l'image de base et l'image compressée se situent en dessous de la sensibilité de notre oeil et il est bien souvent impossible de distinguer les deux images. En effet, l'œil humain parviendrait à discerner beaucoup moins bien des différences de teinte (chrominance) que des différences, mêmes minimes, de clarté (luminance). Cet algorithme exploite donc les failles de notre système visuel pour parvenir, dans ses conditions optimales d'utilisation, à des diminutions de taille d'un ordre de grandeur allant jusqu'à 20 fois (soit 4 fois mieux que GIF) sans dégradation visible à l'oeil. Le temps de compression/décompression est assez important : il faut trouver le bon rapport entre taille de fichier et temps de calcul;
- une dernière possibilité intéressante de ce format est le JPEG progressif. Il s'agit d'afficher l'image non pas en partant du coin supérieur gauche de l'image pour finir, en affichant ligne par ligne, au coin inférieur droit, mais plutôt de donner des vues successives, et chaque fois plus précises, de l'image. L'utilisateur n'a pas à attendre le chargement complet de l'image pour voir s'afficher une première esquisse. Dès le premier passage, il a une idée du sujet. Dans ce système, l'information présente dans la première image est réutilisée pour le second passage, celle de ce dernier pour le troisième et ainsi de suite.

  Un fichier JPEG progressif n'est donc pas beaucoup plus volumineux que son équivalent " *one shot* ". Le temps de calcul est par contre le même pour chaque passage et le temps mis pour calculer une image JPEG en trois passes sera donc trois fois plus important que son homologue " *simple passe* ", mais ce n'est plus guère un inconvénient à la vitesse de nos ordinateurs. *A fortiori*, sur Internet où l'on passe plus de temps à attendre les données transmises sur la ligne téléphonique qu'à attendre que le processeur calcule. L'attente de l'utilisateur regardant s'affiner l'image semblera donc moins longue. Cette méthode s'applique aux images fixes; une méthode dérivée s'applique à la compression des images animées, c'est la norme MPEG.

L'exploitation de fichiers JPEG peut se faire exactement de la même manière que pour les fichiers GIF.

c) format PNG (Portable Network Graphic)

La genèse du format de fichier PNG, initialement appelé PBF pour " *Portable Bitmat Format* " remonte au début des années 90 et à la volonté de créer un format de fichiers capable de remplacer GIF. Il devait être, de l'avis même de ses concepteurs, meilleur, plus petit, plus extensible et surtout gratuit, donc basé sur un algorithme nouveau ou dont les sources auraient été librement disponibles. Un grand débat eut lieu sur plusieurs groupes Usenet pour définir les caractéristiques idéales de ce format et aboutir à l'aspect actuel de PNG :

- ces fichiers supportent 24 bits pour le codage de la couleur (ce que l'on appelle aussi le True Color) ou 16 bits en niveaux de gris (mais il existe aussi une implémentation 8 bits/256 cou-

leurs). Ces spécifications sont donc équivalentes à JPEG;
- ils supportent l'entrelacement des images, de nouveau comme GIF, à la distinction près que l'entrelacement est ici à deux dimensions : on n'entrelace pas des lignes mais des carrés de l'image et la visibilité s'en trouve améliorée dès le premier passage;
- une correction gamma peut être incluse dans le fichier. Les images créées sur PC apparaissent plus claires quand elles sont vues sur un Macintosh et inversément, la correction gamma étant différente dans le réglage des ces deux types de machines. Schématiquement, la correction gamma sert à corriger le contraste des couleurs intermédiaires. C'est un paramètre essentiellement matériel mais quelquefois réglable. En incluant cette correction dans le fichier, l'utilisateur peut adapter son matériel pour se rapprocher le plus possible de ce que le concepteur a voulu créer;
- ils utilisent un canal alpha. Il s'agit d'associer à chaque point de l'image une valeur supplémentaire codée sur 8 bits. Cette valeur est utilisée pour représenter la transparence de ce pixel. On a donc bien 254 niveaux différents de transparence (et non pas 256 puisqu'un niveau correspondra à la transparence complète et un autre à l'opacité complète) et on obtiendra une image dont les points possèderont une transparence plus importante, par exemple, en périphérie qu'au centre. Elle se fondra donc beaucoup mieux sur le dessin donné à l'image se trouvant en dessous. GIF est ici largement dépassé;
- la compression est basée sur une version légèrement modifiée de LZ77, plus efficace parfois que LZW tout en étant gratuit, et elle est sans perte. On peut donc sans crainte compresser et décompresser l'image sans perte d'information ni de qualité visuelle. Enfin, cinq versions de l'algorithme (les compression filters) sont utilisables et, appliquées à différentes parties de l'image selon un certain type d'arrangement des données, permettent des gains de taille appréciables (48 Mb pour une image brute, 36 Mb en compression simple et 120 Ko en compression " *intelligente* " par exemple);
- comme le format PNG a été créé en pensant à l'échange de données graphiques via un réseau, Internet en l'occurrence, certains algorithmes d'autocorrection ont également été implémentés. Il s'agit d'inclure, dans le flot de données constituant l'image, certaines données permettant de reconstruire l'image comme à l'origine même si des erreurs se sont produites pendant la transmission. Il ne s'agit pas de recréer l'image à partir de la moitié des données mais on peut la reconstruire malgré quelques erreurs lors du transfert et sans devoir tout recommencer.

  Techniquement trois types de contrôle d'intégrité sont mis en place : d'abord la " *magic signature* " qui permet de récupérer une image codée en ASCII qui aurait été envoyée en binaire ou inversement; le second est la mise en place d'un CRC (Cyclic Redundancy Check), CRC-32 dans ce cas, qui permet de détecter les erreurs, voire de les corriger, en découpant le flot de données compressées en trames et en leur assignant une valeur calculée pendant la réception du fichier; le troisième enfin, une sorte de CRC des CRC's (le Adler-32 checksum) qui permet un contrôle encore plus efficace des données non compressées;
- il n'existe pas de PNG animés et il ne semble pas prévu d'en faire un jour.

L'implémentation de PNG sur les " *grands browsers* " (Netscape Navigator et Internet Explorer) n'est encore que partielle, alors qu'elle est beaucoup plus aboutie sur de " *petits browsers* " disponibles sous LINUX, le passage par un logiciel de visualisation de fichiers graphiques est donc nécessaire.

1.1.2. Documents textes

Les documents dont il est question ici ont été scannés mais surtout passés par un logiciel de reconnaissance optique de caractères (OCR pour Optical Character Recognition). Cette opération ne peut être

pratiquée, dans l'état actuel de la technique, que par un opérateur humain qui intervient peu ou pas du tout pendant le processus mais doit corriger l'épreuve livrée par un logiciel avouant encore quelques lacunes. Par exemple, la distinction entre 1, chiffre un, et l, lettre L minuscule pose pas mal de problèmes. Ceux-ci devraient être réglés par une reconnaissance incluant le contexte, mais l'industrie du software " *grand public* " n'en est pas encore là. En attendant de tels progrès, seule l'intelligence humaine est capable de donner sa " *vraie* " forme au texte délivré par l'OCR.

A. Fichiers textes ASCII simple

Ce furent les premiers fichiers textes diffusés et de ce fait les plus archaïques techniquement, mais également les plus universellement reconnus, et interprétables, sur toutes les plate formes logicielles, même les plus rudimentaires. A l'origine, le code ASCII (pour American Standard Code for Information Interchange), norme purement américaine, codait les caractères de l'alphabet sur 7 bits, soit 128 possibilités; suffisamment pour coder les lettres minuscules, leurs équivalents majuscules, les chiffres, quelques caractères graphiques (%, & ou encore $) et 32 caractères dits de contrôle qui pouvaient transmettre un ordre et n'étaient pas imprimables, mais pas les alphabets à signes diacritiques. Ce fut possible lorsque le code ASCII passa à 8 bits.

Actuellement, ce format n'est plus guère utilisé. Il présente pourtant certains avantages : des fichiers très petits, donc très rapidement téléchargeables, contenant toute l'information nécessaire pour le traitement, la mise en forme, l'impression et ne demandant pas de configuration informatique musclée. Malgré ces qualités ce format est déconseillé, sauf peut-être pour l'envoi d'articles ou de " newsletters " par email, et est cité à titre historique seulement.

B. Fichiers dans un langage de balisage

Ces langages sont appelés " *markup languages* ". Il s'agit généralement de fichiers textes simples dans lesquels des parties du texte sont entourées d'étiquettes, les tags, qui précisent quelle fonction logique ou hiérarchique joue chaque élément ainsi marqué au sein de la structure du document. Les trois langages que nous allons voir distinguent la structure, le contenu et la forme d'un document. La structure représente le squelette, l'organisation logique du document ainsi que les éléments comme les titres, la division en sections, la zone d'expéditeur, le numéro de révision, ... Les données constituent le contenu qui sera inséré dans les cases définies par la structure. La forme est l'aspect visuel donné au contenu en fonction de son ordre dans la structure : par exemple les titres dans une police plus grande et en gras, les adresses en italique, ...

SGML

La norme ISO 8879, datant de 1986, décrit les règles que doivent suivre les documents **SGML (Standard Generalised Markup Language).** Cette normalisation assure une parfaire compatibilité des documents SGML entre toutes les plates-formes clientes. Cette possibilité d'utiliser le même document sur tout type de matériel, et de pouvoir l'échanger facilement est l'une des grandes forces de ce standard.

SGML est ce que l'on appelle un métalangage, c'est-à-dire une langue possédant son propre vocabulaire et sa propre syntaxe et permettant de créer d'autres langues. Il permet de décrire un document de manière structurée. Typiquement, un document SGML est un fichier texte dans lequel le contenu apparaît entouré de tags. Ces tags précisent la fonction d'un groupe dans la structure. Mais SGML permet aussi de décrire la structure elle-même par ce que l'on appelle une DTD (Document Type Definition) : celle-ci énonce les tags utilisables, leur ordre, leur caractère obligatoire ou non, l'imbrication possible des différents éléments, ... Sans cette DTD, un fichier SGML serait tout a fait inexploitable puisque le logiciel ne verrait qu'un fichier texte rempli de tags.

Un logiciel, le parser, fera le lien entre le document SGML et la DTD pour produire un document tout à fait exploitable.

Evidemment, une telle flexibilité du langage, puisqu'elle permet de créer de nouveaux objets propres à remplir les attentes de l'utilisateur, induit un codage complexe. Une DTD simple peut tenir en quelques lignes, mais un document plus complexe peut prendre aisément plusieurs pages. Cette complexité a éloigné SGML du grand public bien qu'un grand nombre de sociétés ait choisi ce standard pour représenter, archiver et véhiculer le volume important de données qu'elles génèrent. A priori, SGML semble donc être la norme à adopter pour les documents futurs, et ce malgré l'absence à l'heure actuelle d'outils faciles à utiliser. En effet, il faut, pour lire et travailler ces fichiers des logiciels spécifiques comme, par exemple, SoftQuad Panorama.

HTML

HyperText Markup Language fut mis au point, au tout début des années 90, par Tim BERNERS-LEE, du CERN, pour l'envoi rapide et attrayant de résultats de travaux entre chercheurs. En particulier, il voulait que ces documents soient facilement lisibles sur toutes les machines et il souhaitait pouvoir faire des liens hypertextes entre eux. Il se tourna vers SGML et mit au point une DTD adaptée à ses besoins. HTML n'est donc rien d'autre qu'une DTD publique, mise gratuitement à la disposition d'un grand nombre d'utilisateurs. La popularité de HTML est essentiellement due à Internet et principalement au Web, puisqu'il est le langage utilisé pour rédiger les pages d'informations présentes sur le réseau.

HTML propose donc un ensemble de tags, dont le nombre et les fonctions évoluent au fils des révisions du langage (HTML 1.0, puis 2.0 en novembre 1995 et actuellement 4.0), que l'utilisateur peut combiner à son gré pour présenter son information. Il permet des liens hypertextes vers divers types de documents : autres pages Web, images, sons, séquences vidéo, recherches dans une base de données,...

Malheureusement, HTML s'écarte de son grand frère par trois points : d'abord le concept de séparation structure-données-forme disparaît puisque certains tags, plutôt que de signaler qu'un texte est important, demandent que son affichage se fasse de manière particulière, par exemple en gras, mélangeant ainsi la fonction et la forme. Deuxièmement, HTML se révèle sensible à la plate-forme lectrice puisque l'affichage de sa page dépend de la plate-forme qui l'utilisera. Enfin, HTML est limité dans son évolution. Alors que SGML est ouvert, HTML est bloqué dans sa DTD et ses progrès possibles ne concerneront guère que l'implémentation de nouveaux tags - (comme cela se fait dans le Dynamic HTML, DHTML, ou les feuilles de style, CSS pour Cascading Style Sheets, qui sont un essai pour distinguer le contenu de la forme, mais dont les utilisations sont encore rares) - qui ne vont faire que venir alourdir des pages dont le contenu informationnel est déjà bien souvent noyé sous les codes.

Les fichiers HTML sont lisibles avec n'im-porte quel browser Internet; leur création réclame un simple éditeur de texte ou, pour plus de confort, un éditeur HTML.

XML

XML (eXtensible Markup Language), dont la première version normalisée date de février 1998, se pose comme un outil intermédiaire entre SGML, souvent trop lourd et trop complexe, et HTML, très simple à utiliser et à apprendre mais limité dans ses possibilités. XML renoue avec les métalangages : l'utilisateur peut donc de nouveau définir ses propres balises ou utiliser celles déjà prédéfinies. Les impératifs du groupe de travail XML au W3 Consortium étaient les suivants :

- XML devra être directement utilisable sur Internet;
- XML devra supporter une large variété d'applications;
- XML devra être compatible SGML;

- les programmes de traitement des fichiers XML devront être aisés à écrire;
- les caractéristiques optionnelles de XML doivent être réduites au minimum pour assurer un maximum de portabilité;
- les documents XML devront être relativement faciles à lire par un humain;
- la conception de XML devra être achevée rapidement, devra être formelle pour ne laisser que peu de place a l'interprétation et concise pour ne pas retourner vers la complexité de SGML;
- enfin, les documents XML devront être faciles à créer.

Deux outils intéressants en XML sont le XLL (XML Linking Language) qui permet de faire des liens multiples sur Internet entre divers fichiers XML et le XSL (eXtensible Stylesheet Language) qui est l'équivalent des feuilles de style et permet donc de définir plusieurs mises en forme différentes que l'on peut appliquer au même document. Schématiquement, un utilisateur peut donc créer sa propre feuille de style pour afficher comme il le souhaite le contenu structuré des pages qu'il visite. XML intègre également ce que l'on appelle des métadonnées, une évolution des métatags de HTML, qui devraient améliorer l'indexation des documents et aider le travail des moteurs de recherche.
Comme cela était demandé dans les caractéristiques impératives de ces fichiers, ils peuvent être lus par tout browser Internet.

C. Fichiers traitement de texte

Les textes ont été mis en forme à l'aide d'un logiciel commercial de traitement de texte. Cette option présente l'avantage que la plupart des machines sont équipées de ce type de logiciel et que l'utilisateur en connaît déjà plus ou moins le fonctionnement. De plus ces logiciels permettent de faire assez facilement des mises en page évoluées. Mais les multiples inconvénients l'emportent sur les avantages et les traitements de texte sont à déconseiller et d'ailleurs peu utilisés. Il s'agit de logiciels " *propriétaires* ", absolument dépendants de l'éditeur : aucune pérennité n'est assurée pour les fichiers en cas de faillite par exemple, et la compatibilité existe rarement entre les versions successives du logiciel. La visualisation de tels fichiers n'est, en général, possible qu'avec le logiciel " *créateur* "; des conversions sont possibles mais les résultats sont parfois assez différents des fichiers de départ, surtout dans leur mise en forme.

D. Fichiers PostScript et parents

PostScript est un langage de description de page créé par la société ADOBE, connue pour ses logiciels dans le domaine de l'édition électronique aussi bien de textes que d'images. Ce langage a été conçu pour établir un dialogue entre l'ordinateur et l'imprimante, et donner à l'opérateur un contrôle total sur l'aspect visuel de la page. Il est doublé d'un langage de programmation. Très largement utilisé dans le domaine professionnel, l'utilisation de ce langage nécessite généralement un matériel assez coûteux.

Un logiciel de traitement de texte, TEX puis LATEX, avait été créé presque à la même époque pour représenter quantité de symboles et formules mathématiques en utilisant un langage de balisage semblable, dans son principe, à celui à la base de SGML. Il devint très rapidement l'outil " *idéal* " de création de fichiers PostScript. Un troisième type de fichier existe dans la famille PostScript : il s'agit des fichiers DVI (DeVice Independant). En fait, la chaîne de publication partait de fichiers TEX créés sur une machine en utilisant un balisage défini puis étaient transformés en fichiers DVI indépendants de la plateforme de traitement, avant d'être envoyés vers un autre système qui générait le fichier PostScript adapté au matériel utilisé, et se chargeait de l'impression.

Ce type de fichier est également assez peu utilisé car il s'agit aussi d'un format propriétaire qui permet peu de représentation graphique. Les schémas et illustrations doivent y être joints de préférence dans le format EPS (Encapsuled Post-

Script). Son terrain de prédilection demeure l'impression.

Les fichiers PostScript sont essentiellement envoyés sur une imprimante; toutefois l'utilisateur peut ouvrir de tels fichiers avec un viewer PostScript comme, par exemple, GhostScript.

E. Fichiers PDF

Le format de fichier PDF (Portable Document Format), propriété de la société Adobe, présente de nombreux avantages et est en passe de devenir un standard de fait pour l'échange de documents sur les réseaux. En effet :

- les fichiers PDF sont des fichiers images, dérivés des fichiers EPS, qui contiennent des images, des graphiques, des liens hypertextes, de petites séquences sonores ou d'images animées et du texte partiellement éditable. Comme images, les fichiers PDF respectent absolument la présentation, la mise en page donnée par l'auteur, tant à l'affichage qu'à l'impression;
- le format PDF est lisible virtuellement sur toutes les plates-formes informatiques actuelles. ADOBE, par un jeu de dumping, a distribué gratuitement son logiciel de lecture des fichiers PDF, Acrobat Reader. Il n'y a donc plus de problème de matériel non compatible ni de versions différentes du logiciel;
- Acrobat est une famille de logiciels. Le Reader permet, sur une machine cliente, de lire les fichiers. Le Writer permet de composer directement des fichiers PDF avec toutes leurs possibilités. Enfin, le Distiller permet de construire un fichier PDF à partir d'un simple fichier texte ou d'un fichier conçu dans un traitement de texte. Il ne reste plus qu'à lui ajouter, éventuellement, quelques améliorations visuelles. On peut donc, presque sans manipulation de l'auteur, créer un fichier PDF com-plet et utilisable : Distiller se comporte comme une imprimante virtuelle à choisir au moment d'imprimer le docu-ment. Aucune autre manipulation n'est en théorie nécessaire;
- les fichiers générés ne sont pas trop volumineux, ce qui est très important pour l'échange sur de simples lignes téléphoniques. Evidemment, ils sont plus gros que de simples textes ASCII, mais possèdent des mises en page impossibles à réaliser en texte simple. Ils sont beaucoup moins lourds que leurs équivalents éventuels en fichier image;
- PDF est aussi un format propriétaire. S'il semble actuellement s'imposer comme standard, la méfiance reste donc de mise.

Les fichiers PDF ne sont lisibles qu'à l'aide d'Acrobat Reader ou directement dans la fenêtre d'un browser Internet auquel un plug-in, gratuit, a été ajouté.

F. Fichiers RealPage

Ces fichiers sont des " héritiers spirituels " des fichiers PDF. Ils présentent plus ou moins les mêmes caractéristiques, offrent les mêmes avantages mais sont affublés des mêmes inconvénients : il s'agit toujours d'un format de fichier propriétaire, cette fois de la société Catchword Ltd. Cette société, fondée en 1994, propose des solutions " *clé sur porte* " pour la publication électronique de documents et, en particulier, la publication de journaux sur Internet. De fait, cette société assure actuellement l'édition électronique complète de 300 journaux provenant de 24 éditeurs (dont notamment la Royal Society of Chemistry). Cela comprend le scannage/OCR ou le traitement de fichiers textes, la mise au format RealPage, l'indexation pour d'éventuelles recherches en full-text, la mise au point de références actives. Ces dernières sont des sortes de liens hypertextes entre articles d'un même sujet/auteur/domaine/revue ou vers des bases de données spécialisées, des sites web. La firme a conçu, développé et distribue gratuitement le seul logiciel de lecture des fichiers adoptant ce format. Une version Windows est actuellement disponible " *en natif* ", les autres plate-formes doivent passer par une application

Java (offrant les mêmes possibilités mais nécessitant un browser compatible Java 1.1 comme Internet Explorer 4.0 - pour Mac et UNIX - ou Netscape 4.5 - UNIX uniquement). La société peut même proposer un hébergement sur ses propres serveurs (11 actuellement).

La diffusion est basée sur le principe d'un enregistrement - pour un utilisateur seul au moyen d'un " *username* + *password* " ou pour un groupe d'adresses IP- qui implique la réception d'un numéro d'identification (baptisé ici CID pour Catch-Word Identification Number). Le paiement se fait soit au travers de l'abonnement à la version électronique de la revue auprès de l'éditeur, soit suivant le principe du " pay per view ".

Le système est constitué de trois types de logiciels :

a) le browser RealPage, installé chez le client, permet de visualiser et d'imprimer les fichiers rapatriés sur la machine de l'utilisateur;
b) le serveur RealPage, installé au siège de la société, se charge de toute la partie indexation et des références actives. L'accès aux articles est performant : choix du serveur le plus proche, disponibilité des articles 24h/24h;
c) le contrôleur d'accès RealPage, installé chez les éditeurs ou les gestionnaires d'abonnement, leur permet d'enregistrer et de contrôler les paramètres d'un abonnement précis. L'institution cliente ne garde qu'un seul interlocuteur, Catchword, dans une relation qui lui cache ainsi la complexité du système.

Les périodiques ainsi publiés sont disponibles pour au moins trois ans. Après cela, l'éditeur peut soit prolonger son contrat, soit décider de l'interrompre. L'utilisateur n'a donc aucune certitude quand à la pérennité de l'accès.
Le même service s'est ouvert récemment pour d'autres publications : les partitions musicales.

Les fichiers RealPage ne sont lisibles qu'au travers du browser RealPage.

2. Documents prêts pour une publication en format électronique quelconque

Dès la rédaction du document, il est décidé qu'il ne sera pas seulement imprimé sur un support papier, mais aussi publié sous la forme d'une page Web, distribué sur disquette ou intégré à un CD-ROM. Pour que cela soit possible, le codage de l'information doit être " *versatile* " : sa forme peut être changée en quelques manipulations simples, idéalement par programme ou script, donc sans intervention humaine. Si possible, le format choisi ne sera pas la propriété d'une société afin d'éviter tout risque de " *taxation* " lors de l'utilisation ou de la diffusion des documents, mais également pour se mettre a l'abri d'une éventuelle disparition de la société, laissant un document tout à fait figé, techniquement non évolutif. De plus, ce codage devrait être indépendant des machines utilisées pour ne pas vieillir en même temps que le matériel ou les supports.

Actuellement, le langage répondant le mieux à ces exigences est un langage de balisage. On pourrait penser à HTML, très populaire vu son utilisation sur le Web mais ses limites et son absence de normalisation respectée oblige à l'écarter. SGML, grâce à son caractère d'auto-définition (on peut définir le langage et ses évolutions à partir de lui-même), sa grande souplesse et sa normalisation déjà bien avancée semble être le candidat idéal. XML, combinant les avantages de HTML, essentiellement la simplicité, et de SGML serait, aux dires des spécialistes, LE format émergeant. Les outils sont encore rares ou pas assez conviviaux, mais c'est néanmoins le choix conseillé pour les documents futurs.

* * *