::Savoirs LE FIL DES IDÉES

La structuration des documents électroniques

Lieux de savoir, 2. Les mains de l'intellect, Albin Michel, 2011, p. 445-463

Déjean

> Le papier est l'avenir de l'écran.
> Nicolas Taffin.

Depuis ces cinquante dernières années, le livre, et plus généralement le document, semble pris dans un tourbillon tant les innovations dans le domaine des nouvelles technologies sont fréquentes. Certains se sont même demandé si le livre, voire le papier avaient encore un avenir. Si en 1975 on imaginait le bureau sans papier, on sait aujourd'hui que ce mythe du « sans papier » a vécu[1] et que la coexistence et la complémentarité entre support papier et électronique ne font que commencer. Le livre résiste, le document se dématérialise, la bibliothèque devient toile, et les pratiques sociales autour du livre s'adaptent.

Ayant choisi de parler ici du document électronique, nous pouvons d'abord essayer de le définir. Dans cette perspective, Michael Buckland [2] s'est référé dans un article à différents travaux antérieurs consacrés à la notion de document. Parmi toutes les définitions proposées, nous avons retenu celle qui nous semble la plus pertinente ici et qui est proposée par S. R. Ranganathan [3], père de la bibliothéconomie indienne. Elle met en avant plusieurs caractéristiques : un document est l'enregistrement d'un travail sur papier ou sur un autre support, qui permet une manipulation physique facile, un transport dans l'espace et une préservation dans le temps.

L'ajout de la facette électronique au document modifie ces trois caractéristiques : la manipulation devient également électronique et doit être assurée par un ordinateur. Elle gagne en vitesse. Le transport dans l'espace devient quasi immédiat, mais reste parfois dépendant et limité par le format de données et le matériel utilisés. Et nous verrons que sa préservation dans le temps devient problématique.

Les informaticiens se sont emparés du document avec, selon les intérêts en jeu, plusieurs objectifs. Certains veulent accroître l'efficacité de sa manipulation : sa création, son stockage, son édition, sa dissémination. Le document électronique devient alors donnée à traiter automatiquement, et l'on oublie parfois sa spécificité par rapport aux autres types de données. D'autres s'intéressent à sa représentation visuelle, plus traditionnelle, et à son transfert vers le numérique : des équations remplacent les fontes en plomb, le document électronique est le produit de calculs.

C'est l'histoire rapide de ces deux approches technologiques que nous allons retracer. Quel que soit l'objectif des travaux, objet de traitement ou représentation visuelle, on a rapidement identifié plusieurs besoins, entre autres l'interopérabilité, c'est-à-dire l'indépendance des données par rapport au support matériel, mais aussi la pérennité dans le temps, une des facettes qui caractérisent le document selon Ranganathan. Si la première exigence semble aujourd'hui en voie de satisfaction grâce à la mise au point de standards ouverts, la préservation des documents électroniques, qui s'appuie aussi sur ces standards, pose encore un grand nombre de questions.

Le document électronique a quarante ans, et les technologies dont il dépend sont encore en évolution. Tous les dix ans, de nouvelles possibilités apparaissent. Nous verrons donc comment les acteurs du document, scripteurs, lecteurs, éditeurs, réagissent et s'adaptent face à ces changements.

Dans ce texte, nous nous sommes permis de reprendre certains titres de sections de l'ouvrage de Lucien Febvre et Henri-Jean Martin [4] ; le lecteur pourra s'y référer pour approfondir le parallèle entre document électronique et document « traditionnel ».

Les difficultés techniques et leurs solutions

Paraphrasant Henri-Jean Martin, on affirmera que l'on ne peut isoler l'histoire du document électronique de celle des inventions qui l'ont précédé ou accompagné. L'histoire de l'ordinateur, bien qu'assez courte, serait trop longue à raconter ici. Mais il est évident que les capacités de calcul et de stockage de l'information eurent et ont toujours un impact capital sur les utilisateurs et les utilisations. Si les premiers ordinateurs n'avaient pas la capacité actuelle de calcul et de stockage, ces limitations tombèrent vite, mais eurent parfois une incidence sur la manière de représenter un document[5]. L'évolution

des périphériques fut similaire, et l'apparition de l'ordinateur individuel dans les années 1980 assura sa large dissémination et son succès. Cette évolution technologique continue tant au niveau de la puissance de calcul et de la qualité des périphériques qu'au niveau du coût d'achat et d'utilisation. Mais la puissance de calcul ne suffit pas : il faut pouvoir représenter ces documents et les manipuler.

Selon Charles F. Goldfarb [6], la représentation informatique du texte dans un système de traitement intégré doit permettre de partager l'information entre différents utilisateurs quelle que soit l'application. C'est à ce problème de l'interopérabilité que répondent en partie aujourd'hui les normes et standards mis au point au cours des dernières décennies. Nous allons faire rapidement leur historique en allant du caractère au document qui, sous l'effet des derniers standards comme XML, perd de sa spécificité et se transforme aujourd'hui en donnée.

La fabrication des caractères

'Voici où nous en sommes. Aujourd'hui, grâce à votre ordinateur, vous pouvez composer n'importe quelle langue – pourvu que ce soit de l'anglais[7].'

Commençons par le b.a.ba du document : le caractère et son codage. Le codage exprime la manière dont l'information est structurée au sein du fichier de façon à pouvoir être conservée, transmise et échangée. Une table de codage des caractères associe à un caractère une représentation numérique, et c'est cette représentation qui est manipulée par l'ordinateur. La première table de codage standardisée de caractères proposée par l'American National Standards Institute en 1963 comportait 128 caractères, dont seulement 63 visibles[8], les autres étant des caractères de contrôle ou n'étant pas employés. Elle héritait de toute l'histoire du codage de l'information et des télécommunications[9]. Ce codage et ses versions suivantes furent utilisés par tous les constructeurs d'ordinateurs américains à l'exception notable d'IBM qui utilisa son propre codage, EBCDIC. L'inadéquation évidente entre cette table de codage sur 7 bits et les différents systèmes d'écriture suscite l'apparition de nombreuses tables de codage pour les différents systèmes d'écriture alphabétiques, syllabiques et idéographiques en usage dans les pays concernés. La prolifération et l'incompatibilité de ces différents systèmes étant un frein à l'échange de documents et à la manipulation de documents multilingues, cette situation conduisit à la création dans les années 1990 du consortium Unicode, comprenant les principaux acteurs industriels du domaine, et à la publication en 1991 d'un standard éponyme. L'objectif de ce standard Unicode fut de créer un codage pour plus de 65 000 caractères et symboles, mathématiques et chimiques par exemple, permettant ainsi le codage de documents multilingues. La dernière version d'Unicode, Unicode 5.0, permet de

représenter tous les systèmes d'écriture passés ou présents et est intégrée dans de nombreux standards actuels, comme XML. Si cette solution technologique semble résoudre beaucoup de difficultés[10], le problème des fonds numériques existants qui n'utilisent pas Unicode se pose : actuellement de nombreuses solutions informatiques ne se servent pas de ce standard. Et l'intégration d'un tel standard dans un système informatique existant nécessite un tel effort qu'il est parfois plus simple de changer de système. On comprend alors l'inquiétude des entreprises face à ce choix. Le propos d'Ari Davidow cité au début de cette section est donc d'actualité pour quelques années encore.

Représenter le texte : balisage procédural et balisage descriptif

Du caractère passons maintenant au texte. Le développement des premiers éditeurs[11] de texte ne vint pas d'une demande de l'édition traditionnelle, le coût d'un ordinateur étant alors prohibitif, mais du besoin toujours grandissant des informaticiens soucieux d'imprimer leurs programmes et de créer leur documentation. Le problème du formatage de texte se posa alors : comment ajouter une mise en page et des effets typographiques à ces documents ?

Les premiers formalismes, comme RUNOFF, correspondaient à un flot textuel dans lequel des instructions de formatage étaient insérées. Selon le manuel d'aide du langage de formatage ROFF, ces annotations suivaient les habitudes des typographes de l'âge pré-informatique, où des lignes commençant par un point étaient utilisées dans les manuscrits pour noter des demandes de formatage. Ce type de formatage est appelé *procédural :* les annotations insérées spécifient comment le texte doit être formaté, par exemple concernant le centrage et la taille des caractères. Ce type de langage commença à soulever quelques critiques : il était souvent lié à un périphérique particulier comme une imprimante, les opérations de mise à jour du document étaient parfois compliquées, les auteurs devaient consacrer du temps à la mise en page. Et se posa déjà la question de la pérennité du langage : celle-ci dépendait essentiellement du constructeur de périphériques. Les éditeurs et les bibliothèques qui commençaient à utiliser ces langages se trouvèrent vite confrontés à ces problèmes.

Vers la fin des années 1960 fut créé sous l'impulsion du Graphic Communication Association (GCA), le comité GenCode, pour Generic Coding, dont le but était de mener une réflexion sur la séparation du contenu du document et de son formatage, et par là même de permettre un meilleur échange des informations. La paternité de cette idée est attribuée à William W. Tunnicliffe, membre du GCA, qui l'exposa en 1967. Le comité GenCode arriva à la conclusion qu'il était impossible de concevoir un ensemble de balises qui pût satisfaire tous les types de documents. Ils formalisèrent leurs conclusions sous la

forme de concepts, connus sous le nom de GenCode Concepts, qui sont les suivants :

- des codages génériques différents sont nécessaires selon les types de documents. Autrement dit, on ne peut concevoir un jeu de balises « universel » couvrant tous les types de documents ;
- les petits documents peuvent être incorporés comme éléments de documents plus grands. On trouve ici la notion de structure hiérarchique d'arbre.

Ces concepts jouèrent un rôle important dans le développement du standard SGML que nous allons aborder maintenant.

On doit la première réalisation de ces concepts à Charles F. Goldfarb, un juriste de formation travaillant pour IBM sur un projet de recherche d'information dans le domaine juridique. Avec Edward Mosher et Raymond Lorie, il développa le langage GML (Generalized Mark-up Language). Ce travail de Goldfarb servit de base à la mise au point de SGML (Standardized Generalized Markup Language), qui devint une norme de l'organisation internationale de normalisation *(international standard organisation)* en 1986. En quelques lignes[12], un document balisé avec SGML est organisé hiérarchiquement en éléments identifiés par des balises le décrivant. Par exemple, ce texte peut être balisé avec des balises *chapitre*, *section*, *paragraphe*, mais aussi *note*, ou *auteur*. On parle alors de *document structuré*. Ce balisage qui reflète la structure sémantique du document ne doit faire référence à aucune information de formatage. Les informations de formatage sont contenues dans une feuille de styles[13], où le formatage de chaque balise est précisé. De plus, un modèle de document[14] est associé à chaque document, permettant ainsi sa validation. Le langage SGML est en fait un métalangage où l'ensemble des balises n'est pas défini *a priori*, et il doit être vu comme une technique qui permet de spécifier un jeu de balises et leur organisation hiérarchique, pour un type de document particulier.

Un descendant direct de SGML est le langage HTML développé par Tim Berners Lee, chercheur au CERN, qui, s'il permet de publier des informations sur le Web, fut critiqué par la communauté SGML pour son non-respect de certains principes de base. En particulier, la présence de balises de formatage n'assure plus la séparation entre contenu et forme. Mais sa facilité d'apprentissage et d'utilisation est telle que son succès a été immédiat et mondial. La communauté SGML n'abandonna pas pour autant le combat et mena une réflexion sur l'utilisation de SGML, dans sa version pure, pour le Web. Des difficultés techniques liées à SGML, comme sa complexité par rapport à HTML, ou le problème de la validation de documents vis-à-vis de son modèle, amenèrent ce groupe, comprenant en particulier Jon Bosak, Tim Bray et James Clark, à proposer une version simplifiée de SGML : XML (eXtended Mark-up Language), qui devint une norme du W3C[15]

en 1998. Capitalisant sur l'expérience de SGML, XML connut un succès rapide dans la communauté scientifique, et il est aujourd'hui fréquemment utilisé comme format de codage et d'échange de documents, et plus généralement de tout type de données. Se fondant sur des standards internationaux comme Unicode, ayant une spécification bien décrite et publique, tout un ensemble d'outils, comme des validateurs, furent vite à la disposition de la communauté. Des langages spécifiques à un besoin donné furent également rapidement développés, tels que SMIL pour le multimédia, RDF pour les ressources et MathML pour les formules mathématiques.

De manière similaire à SGML, le document XML ne contient, théoriquement, aucune information sur le formatage de ses données, lequel repose sur une feuille de styles écrite en XSL (eXtended StyleSheet Language). Cette feuille de styles décrit comment le contenu de chaque balise doit être formaté. Par exemple, le titre d'une section est affiché avec telle fonte et avec telle couleur. Cette possibilité de séparer forme et contenu est aujourd'hui appréciée par les éditeurs lorsqu'ils doivent afficher un message sur une pluralité de supports : une page Web, une version papier, l'écran d'un téléphone portable.

Le typographe moderne

Retour aux années 1960. Logiciels et matériel étaient alors très liés : le constructeur de matériel vendait aussi les logiciels nécessaires pour utiliser ce dernier, et ces logiciels ne fonctionnaient souvent que sur ce matériel. Cette dépendance, si elle n'interdisait pas formellement l'échange de données entre systèmes de différents constructeurs, l'entravait fortement. Ainsi un document dans un format donné ne pouvait-il être imprimé que sur une imprimante particulière. Dans les années 1970, John Warnock, un chercheur travaillant chez Xerox, développa un langage de description de page, appelé Interpress[16], indépendant de l'imprimante utilisée[17]. Warnock et son manager, Chuck Geschke, qui avaient échoué à convaincre leur compagnie de commercialiser ce langage, décidèrent de créer une nouvelle société pour le mettre sur le marché. Adobe Systems Inc. était né. Le nom commercial du langage fut PostScript, et sa première version fut commercialisée en 1984. L'avantage que ce langage offrait était une indépendance vis-à-vis du matériel, et donc d'un fabricant donné, et des spécifications publiques. Ainsi un fabricant pouvait-il facilement concevoir et intégrer dans son imprimante un logiciel, le pilote, qui supportait ce langage.

Ce que firent la société Apple et son président, Steve Jobs, si convaincus par ce langage qu'ils l'intégrèrent dans leur imprimante LaserWriter. Conjointement, une petite compagnie, Aldus, créa le

premier logiciel de publication assistée par ordinateur, PageMaker, fonctionnant sur l'ordinateur Mac et utilisant une LaserWriter. Ce produit fut adopté d'abord par la compagnie Linotype, et le monde de l'édition graphique suivit, faisant de PostScript le langage employé dans le monde de la pré-impression. Et par là même, la fortune d'Adobe, Apple et Aldus. Adobe a depuis lors développé un successeur à PostScript, le fameux Portable Document Format, ou PDF. De la même manière que l'on peut considérer XML comme une version simplifiée de SGML, PDF est un langage de description de page plus simple que PostScript, mais établi sur un principe similaire : préserver le contenu et la mise en forme lors de ces échanges.

Le document électronique : un objet calculé

La qualité des incunables, les premiers livres imprimés, rivalisait avec la qualité des manuscrits, et l'exemplaire de la Bible de Gutenberg conservé à Mons est toujours considéré comme le plus bel imprimé du monde. Cela ne fut pas le cas pour les premiers documents électroniques produits et imprimés. Le lecteur curieux trouvera dans les actes des conférences sur l' *Electronic Publishing* publiés par Cambridge University Press certains articles possédant des colophons mentionnant les outils de rédaction et le matériel d'impression utilisés. Il se fera alors une idée de la qualité obtenue dans les années 1980.

Cette qualité était donc loin d'être comparable à celle de l'édition classique. Mais les progrès furent rapides et ils permirent aussi un renouveau dans certains domaines comme la création d'alphabets numériques. Le format électronique a également ouvert le document au multimédia.

La qualité d'un document imprimé dépend bien sûr de la qualité d'impression de l'imprimante utilisée. Cette qualité d'impression atteignit un niveau satisfaisant dans les années 1980 avec deux technologies toujours d'actualité : l'impression « jet d'encre » et l'impression « laser », cette dernière, couplée à un système d'édition numérique, commençant à concurrencer la technologie offset pour certaines demandes.

Mais la qualité d'un document dépend aussi de sa typographie et de sa mise en page. Ici, deux étalons de comparaison peuvent être cités, à savoir, d'une part, les systèmes d'impression de la presse traditionnelle – dans ce cas, la comparaison fut longtemps défavorable – et, d'autre part, la machine à écrire et sa fonte non proportionnelle – dans ce cas, le document numérique la supplanta rapidement.

Nous allons maintenant voir, en commençant par les fontes

numériques, comment le rendu visuel d'un document est construit. Nous renvoyons le lecteur curieux à l'excellent ouvrage de Richard Rubinstein [18] pour une introduction détaillée à la typographie numérique.

Les caractères : du plomb à l'équation

Le standard Unicode distingue le *caractère* du *glyphe :* le terme *caractère* est utilisé au sens de la représentation abstraite d'une lettre, par exemple le caractère « a » renvoie à la première lettre de notre alphabet. Le *glyphe* est une représentation concrète du caractère, comme la réalisation d'une lettre associée à une fonte. Si Unicode offre une solution technique pour la représentation des caractères, il reste à représenter les glyphes. Les premières fontes numériques furent développées pour les besoins de la photocomposition. Les premiers essais réalisèrent un glyphe sur forme d'image matricielle appelée *bitmap*. Pour chaque taille voulue, il faut, avec ce formalisme, dessiner un alphabet spécifique. Cette solution fut remplacée dans les années 1980 par une représentation vectorielle des glyphes. Une des idées[19] d'Adobe lors de la mise au point de PostScript fut d'utiliser un formalisme mathématique relativement nouveau : les courbes de Bézier, du nom de l'ingénieur qui les mit au point. Le glyphe est alors représenté par un ensemble de lignes et de courbes qui marquent son contour. Ce formalisme vectoriel a l'avantage de fournir des caractères en grande taille sans effet de marche et d'offrir une version indépendante de la résolution des écrans et des imprimantes.

L'inconvénient est que ce formalisme requiert une plus grande puissance de calcul et que le rendu d'un glyphe, c'est-à-dire la conversion de la représentation vectorielle vers une représentation matricielle – un écran étant une matrice de pixels –, exige un savoir-faire particulier et même protégé par des brevets : les algorithmes d'optimisation ou « *hinting* ». Si l'effet de marche a disparu pour les grands corps, c'est-à-dire de grande taille, un effet de *bavage* peut se produire pour les petits corps et pour les périphériques de basse résolution.

Ainsi la qualité d'une fonte vectorielle ne se mesure-t-elle plus seulement à la qualité de ses glyphes, mais aussi à la qualité de ses algorithmes d'optimisation, véritable savoir-faire des typographes numériques[20]. On note que la plupart des fontes utilisent des algorithmes d'optimisation par défaut. Les nouveaux formats de fontes comme OpenType ajoutent aux caractères décrits un ensemble de règles de transformation et de positionnement, rendant les fontes de plus en plus intelligentes, ce qui explique aussi pourquoi les fontes numériques sont considérées, du point de vue de la propriété intellectuelle aux États-Unis, comme des logiciels, et donc soumis au

copyright, alors que les fontes traditionnelles en plomb ne peuvent être protégées par un *copyright*, à l'exception de leur nom qui seul peut l'être.

Du compositeur à l'algorithme

> Et le compositeur ? Il a disparu. Reviendra-t-il jamais ?
> Fernand Baudin.

Une fois la représentation des glyphes assurée, il faut les mettre en page, c'est-à-dire les positionner dans la page. Le principal problème est de diviser en lignes un flot de texte, généralement un paragraphe. Au temps de l'imprimerie « à plomb », le compositeur se chargeait de mettre les caractères sur un composteur et de régler la justification en insérant des espaces[21]. Comme Don Knuth le relève, il ne semble pas exister de descriptions précises de méthodes pour effectuer cette mise en ligne[22]. Une formalisation rigoureuse du problème de l'insertion de saut de ligne attendra les années 1950, date à laquelle est déposé le premier brevet décrivant en détail une telle méthode. Le problème de la segmentation en lignes n'est pas le seul à devoir être résolu : les traitements de texte doivent aussi effectuer une mise en page de certains éléments flottants comme les figures et les tableaux. Là aussi des algorithmes ont été mis au point ; et, si aucune solution vraiment convenable et robuste n'a été trouvée, la recherche continue dans ce domaine.

Chaque éditeur de logiciel de traitement de texte ou de PAO possède son ou ses algorithmes, et un utilisateur peut comparer les différentes mises en ligne et mises en page d'un même texte. Les logiciels de PAO obtiennent généralement un résultat de meilleure qualité pour la mise en ligne, mais la mise en page, c'est-à-dire le placement de chaque élément du document dans une page, doit être faite à la main. Le traitement de texte peut du reste s'en charger automatiquement.

Si les outils de composition de texte, comme Adobe InDesign et QuarkXPress, atteignent une qualité suffisante pour le monde de l'édition, les outils utilisés pour la composition de texte sur écran, donc pour composer des pages Web, n'ont pas encore atteint cette qualité. Celle-ci devrait être atteinte avec la prochaine génération de feuille de styles en cascade, CSS 3, développée par le W3C. C'est peut-être grâce à ces feuilles de styles que le compositeur reviendra.

Si l'outil électronique n'a pas été à l'origine d'innovations typographiques majeures, comme le disait Bodoni à son époque, le manuscrit sert toujours de référence. Mais certains objets difficiles à produire avec les moyens traditionnels, comme les équations mathématiques, sont devenus communs dans les documents électroniques, grâce à un langage comme LaTeX. Certaines mises en page, celle des notes marginales (par exemple), restent difficiles quels

que soient les moyens utilisés.

Si certains problèmes subsistent, un traitement de texte comme celui développé par le consortium OpenOffice étend continuellement ses fonctionnalités typographiques. Confirmant le propos de F. Baudin, le scripteur arrivera bientôt, grâce à l'outil informatique, à mettre en œuvre tout le patrimoine graphique de l'humanité.

L'illustration

De même que le mélange du texte et de l'illustration ne posa pas de problème technique au début de l'imprimerie[23], le document électronique intégra assez rapidement la notion d'objet documentaire et étendit ces types d'objets aux autres médias connus : le son et la vidéo.

Autre aspect visuel du document : la couleur. Il faudra quelques années pour que la couleur fasse son apparition dans le document et que les écrans couleur se développent. Le travail de normalisation de la Commission internationale de l'éclairage (CIE), créée au début du xx^{e} siècle, va grandement faciliter l'intégration des couleurs, puisque les modèles utilisés actuellement avaient déjà été mis au point, aussi bien le modèle additif pour les écrans, où l'on mélange trois couleurs primaires pour obtenir une autre couleur, que le modèle soustractif pour les imprimantes, où l'on utilise des couleurs pour filtrer certaines longueurs d'onde et obtenir ainsi une nouvelle couleur. La transposition de ces modèles fut donc rapide. Si la couleur a envahi nos écrans, elle reste encore discrète dans le document imprimé, le coût d'une impression en couleur étant encore largement supérieur au coût d'une impression en noir et blanc. Là aussi, comme cela s'est produit pour les autres composants technologiques, une baisse de ces coûts devrait entraîner une augmentation de l'usage de la couleur dans le document imprimé.

Le petit monde du livre

Après avoir décrit les évolutions technologiques, nous allons maintenant nous intéresser à trois acteurs qui gravitent autour du document : le scripteur, le lecteur, l'éditeur. Ou, plutôt, les scripteurs, les lecteurs, les éditeurs. Car chacune de ces activités peut prendre des aspects très différents selon son contexte.

Le nouveau scripteur

La place manque ici pour décrire le travail du scripteur dans les différentes situations de rédaction, qu'elles soient ou non professionnelles : dactylographe, auteur de romans, rédacteur technique, employé rédigeant un document de travail, etc. Nous nous

focaliserons donc sur un scripteur professionnel particulier : le rédacteur technique. Il est aujourd'hui le plus fortement exposé aux nouvelles innovations. Une partie de son travail consiste à mettre en forme la documentation rédigée. Cette documentation qui intègre de nos jours un système de gestion documentaire est, grâce à ce système, reliée à tout un ensemble de données documentaires. Afin que le traitement automatique de cette documentation soit plus facile, un format structuré comme XML ou SGML est utilisé[24]. Ainsi ce document pourra-t-il être adapté et matérialisé sur plusieurs supports, comme le papier ou les écrans, où des vérifications de la conformité avec une norme pourront être faites. Si ce format permet une automatisation de certains traitements, il donne un surplus de travail au rédacteur : celui-ci doit non seulement rédiger et saisir le texte, mais aussi saisir le balisage du texte. Cette opération, assez nouvelle pour la plupart des rédacteurs, est très laborieuse, les balises pouvant correspondre à un mot du texte. Des éditeurs apparaissent pour assister, pré-contraindre pour certains, ce travail de saisie. Mais les interfaces actuellement proposées ne semblent pas répondre de manière satisfaisante aux exigences nouvelles des documents structurés, en particulier lorsque le modèle de document est très complexe, comme c'est le cas dans le domaine aéronautique ou automobile. En outre ces outils devenant de plus en plus puissants, ils nécessitent une formation des utilisateurs[25]. Les interfaces WYSIWYG[26], conçues pour faciliter le travail, ont en réalité caché la véritable complexité que ces outils renferment et que la rédaction structurée semble remettre au jour.

Le lecteur ou les lecteurs qui deviennent bibliothécaires

Après le scripteur, tournons-nous vers le lecteur. On trouve chez David Levy [27] un propos intéressant sur les habitudes que prend le lecteur confronté aux nouvelles technologies : « Les changements technologiques et la nature de la vie moderne pourraient mettre fin à la lecture approfondie... Ce n'est pas que le livre ait disparu, mais plutôt que les conditions culturelles pour la lecture approfondie sont en voie de disparition rapide. »

Ainsi le lecteur contemporain survolerait ses documents et ne pourrait plus lire en profondeur. Ce propos, loin d'être nouveau, a déjà été tenu à l'époque de la scolastique, lors de l'apparition d'outils facilitant cette lecture rapide. Car déjà à cette époque, les étudiants ne pouvaient plus « tout » lire. La faute en revient aujourd'hui à la toile, au Web. Mais faut-il comparer la navigation sur la toile avec la lecture ? Selon Michel Melot, *l'ordinateur ne concurrence pas le livre mais la bibliothèque* . Nous passerions donc maintenant notre temps à vagabonder dans les couloirs d'une bibliothèque en feuilletant rapidement des livres et en oubliant de les lire ? Mais qui n'a jamais passé plusieurs heures dans une bibliothèque pour finalement n'emprunter qu'un livre ?

Pour lire, l'ordinateur ou plutôt l'écran n'est pas encore un bon outil.

Cela n'est pas seulement dû à la mauvaise résolution des écrans, encore loin de la résolution d'une imprimante, mais aussi à la mauvaise qualité des outils de composition de textes et au manque de compétence des rédacteurs. Néanmoins, ces trois éléments n'empêchent pas le développement de la lecture sur écran pour certains types de documents comme les quotidiens. Ce mouvement devrait être amplifié par les progrès technologiques, comme le montrent les nouvelles générations d'écrans très haute résolution, mais aussi par un meilleur support du côté des outils de composition de texte.

Toutefois la préférence pour le papier que manifestent la plupart des lecteurs vient aussi des avantages que ce dernier procure. Ces avantages sont bien connus[28] : la navigation dans un document papier est plus rapide que sur écran, le travail avec plusieurs documents papier est plus facile, ainsi que le travail collaboratif ou la discussion autour d'un document papier. Citons enfin la prise de notes, sur le document même ou bien sur un document séparé. Nous devons ajouter que ces observations concernent des activités de lecture dans un cadre professionnel et que, comme les auteurs de ces travaux le mentionnent eux-mêmes, il n'y a pas un lecteur mais des lecteurs, selon l'activité dans laquelle se situe la lecture.

Pour finir, revenons au livre électronique. Comme certains l'ont remarqué, ce livre électronique est davantage une sorte de bibliothèque, puisqu'il peut contenir des milliers de livres. À la fin de sa célèbre nouvelle, Borges citait dans une note une amie qui lui faisait remarquer que la bibliothèque de Babel était inutile puisqu'un seul volume contenant un nombre infini de feuilles infiniment minces suffisait[29]. Si ce livre va bientôt voir le jour grâce à l'électronique, l'amie de Borges se trompe si elle pense qu'une bibliothèque peut se réduire à un ensemble de feuilles. Le rôle de la bibliothèque, et de ses bibliothécaires, est d'organiser cet ensemble afin de permettre à l'utilisateur de le parcourir et de trouver les livres qui l'intéressent. Ainsi le lecteur d'hier, face à la toile ou à son livre électronique, n'est-il pas obligé de se transformer aussi en bibliothécaire s'il ne veut pas prendre le risque de se perdre dans la quantité d'information à laquelle il a accès ?

Du libraire en ligne à l'imprimeur à la demande

Les acteurs traditionnels du document, éditeurs, imprimeurs et libraires semblent être ceux pour lesquels les changements apportés par les nouvelles technologies sont les plus visibles et les plus rapides. Les librairies électroniques se sont installées sur la toile, comme celle qui est appelée aujourd'hui la plus grande librairie du monde, Amazon.com. Outre le fait de permettre une recherche dans un très large catalogue, cette entreprise offre maintenant des services à ses clients internautes qui tendent à imiter ceux de la librairie

traditionnelle et ses libraires, comme un service de recommandation qui est établi à partir des ouvrages que d'autres clients ont achetés, clients ayant déjà commandé l'ouvrage que vous regardez. C'est une manière automatique d'offrir à ses visiteurs des conseils que les libraires peuvent fournir de vive voix. Depuis peu, l'internaute a la possibilité de feuilleter certaines pages du livre qu'il consulte comme lorsqu'il se promène dans les rayons d'une librairie traditionnelle. Ainsi, de même que le manuscrit sert de référence aux documents imprimés, de même la librairie traditionnelle semble-t-elle servir de référence aux librairies virtuelles. Le développement des mondes virtuels en trois dimensions ne fera peut-être que renforcer ce parallèle.

Du côté des éditeurs, un nouveau concept est apparu : *l'impression à la demande*. Cette impression à la demande, qui consiste à n'imprimer un document que lorsqu'il est demandé par un lecteur, est maintenant possible à faible coût grâce à la qualité des presses numériques qui concurrencent de plus en plus la qualité de l'offset, technique de référence dans le monde de l'impression. Certaines compagnies offrent déjà ce service aux auteurs à qui il suffit de télécharger leurs documents en format électronique.

Si les éditeurs peuvent être court-circuités par les auteurs, ils ne s'adaptent pas moins et proposent, grâce à la toile, de nouveaux services. Le dernier-né du mariage entre édition traditionnelle et Internet s'appelle le *digital object identifier*, ou DOI. Le DOI peut être vu comme un mélange du traditionnel code ISBN et d'un hyperlien. Comme identifieur, il permet de référer de manière unique à un document numérique ou à une partie de ce document. Comme hyperlien *de nouvelle génération*, il permet grâce à un système de gestion centralisé une mise à jour automatique de ce lien, et évite ainsi le problème de l'hyperlien cassé.

La préservation du document électronique

De la première page Web créée au CERN en 1990, il ne reste que la version de 1992, dont une copie se trouve hébergée par le consortium W3C[30]. Aucune copie d'écran de la première page n'a été faite, les plus anciennes datant de 1992 [31]. Si la préservation de la première page Web a, peut-être, seulement une importance historique, le problème se pose de manière plus cruciale pour certains documents électroniques dont la préservation est imposée par des textes législatifs et réglementaires. La recommandation[32] du 1er décembre 2005, rédigée par un groupe de travail du forum des droits sur l'Internet et de la Mission Économie Numérique, intitulée « La conservation électronique des documents » offre une bonne introduction au problème, d'un point de vue juridique, et des

réflexions sur les modalités que doit remplir une conservation électronique fondée sur trois critères : la lisibilité du document, la stabilité du contenu informationnel, et la traçabilité des opérations sur le document. Des réflexions similaires sont actuellement menées à tous les niveaux institutionnels.

Que ce soit dans un système fermé ou dans un système ouvert comme le Web, les responsables des politiques de conservation doivent se poser les questions suivantes : que conserver ? Pour quelle durée ? Et comment assurer cette conservation ? Au regard de la première question, la loi nous rappelle que la réponse ne peut être : tout. Les données personnelles sont en effet protégées par le législateur[33], ce qui peut poser problème lors de l'archivage de données visibles sur la toile. Les durées de conservation sont aussi régulées par la loi, même si une harmonisation au niveau européen est souhaitable. Les pistes de réflexion proposées dans la recommandation sur les modalités techniques de la conservation s'appuient sur un ensemble de bonnes pratiques, sur la notion de métadonnées, informations associées au document et le décrivant, et sur l'utilisation de standards ouverts, c'est-à-dire des formats de données dont les spécifications sont publiques, et gérés par un organisme de normalisation. Ici réapparaissent XML, Unicode et consorts. S'il va de soi que l'utilisation de ces formats n'offre aucune garantie à long terme, ils augmentent considérablement la pérennité des données à court terme.

Un dernier problème non moins important et bien connu dans le monde de l'archivage se pose avec les supports des documents électroniques, car les supports magnétiques sont aussi fragiles que le papyrus.

Conclusion

La technique progresse et nous offre aujourd'hui de nombreuses possibilités : sans sortir de chez nous, nous pouvons être auteur, typographe, maquettiste, éditeur, distributeur de nos œuvres. Cette collusion des métiers n'est pas nouvelle, comme le relèvent Febvre et Martin : « On ne s'étonnera pas si, de tout temps, des écrivains se firent imprimeurs et libraires[34]. » De nos jours, tous les écrivains ont cette possibilité. Une seule chose semble nous manquer : le temps. Et peut-être une formation de base pour chacun de ces métiers.

Notes

1. Sellen et Harper, 2002.

2. Buckland, 1998.
3. Ranganathan, 1963.
4. Febvre et Martin, 1999.
5. Les premiers documents SGML étaient structurés avec des balises de 1 caractère pour ne pas trop prendre d'espace en mémoire.
6. Goldfarb, 1991.
7. Ari Davidow, cité dans Baudin, 1994, p. 419.
8. Pour sa première Bible, Gutenberg disposa de plus de trois cents symboles conçus par Peter Schöffer.
9. Comme le CCITT, Comité consultatif international télégraphique et téléphonique.
10. J. A ndré, 2002.
11. On trouvera dans Haigh, 2006, et Bergin, 2006, une histoire détaillée des outils de traitement de texte.
12. Le lecteur peut se référer à GOLDFARB, 1991, pour une présentation complète de SGML.
13. Appelée DSSSL : Document Style Semantics and Specification Language.
14. Document Type Definition.
15. Le World Wide Web Consortium, ou W3C, est un consortium fondé en octobre 1994 pour promouvoir la compatibilité des technologies du World Wide Web.
16. Développé par Bob Sproull et Bulter Lampson à partir de Press format et JaM.
17. Si celle-ci contient un logiciel, le pilote, qui lui permet d'interpréter ce langage.
18. Rubinstein, 1988.
19. Le système Metafont de Donald Knuth, développé dans les années 1970, est lui aussi établi sur une représentation vectorielle des caractères. Voir Knuth, 1999.
20. Une partie de ce savoir était auparavant contenu dans des tables de crénages.
21. Selon Fernand Baudin, la formation d'un compositeur durait cinq ans.
22. On trouvera néanmoins chez quelques auteurs certaines listes de règles, comme celle de Pierre Simon Fournier pour les correcteurs en typographie, donnée dans Baudin, 2006, p. 273.
23. Febvre et Martin, 1999, p. 133. Propos nuancé dans MARTIN, 1988, p. 219-220, où il est dit que la qualité des illustrations fut au début fruste et que leur insertion provoquait une désorganisation de la mise en page.
24. Bien que de nombreux rédacteurs utilisent encore de simples outils de bureautique.
25. B. André, 2006.
26. *What You See Is What You Get :* rendu graphique du document que l'on saisit. Ce que l'on voit à l'écran est une image assez fidèle de la version imprimée.
27. Levy, 2001, p. 106

28. Sellen et Harper, 2002 ; Marshall, 2003.

29. Borges, 1993, p. 498.

30. [www.w3.org/History/19921103-hypertext/ypertext/WWW/TheProject.html]

31. On trouve une copie d'écran à cette adresse : [info.cern.ch/NextBrowser.html].

32. [www.foruminternet.org/ telechargement/ documents/ reco-archivage-20051201.pdf]

33. Le principe du droit à l'oubli est consacré par l'article 6-5° de la loi du 6 janvier 1978.

34. Febvre et Martin, 1999, p. 210.

Bibliographie

- André, 2002 : Jacques André, *Unicode, écriture du monde ?*, document numérique, 6 (3-4).
- André, 2006 : Bernard André, *Utilisation des progiciels : identification d'obstacles et stratégies de formation*, thèse de doctorat, École normale supérieure de Cachan.
- Baudin, 1994 : Fernand Baudin, *L'Effet Gutenberg*, Paris.
- Bergin, 2006 : Thomas J. Bergin, « The Origins of Word Processing Software for Personal Computers : 1976-1985 », *IEEE Annals of the History of Computing*, 28 (4), p. 32-46.
- Borges, 1993 : Jorge Luis Borges, « La Bibliothèque de Babel », in *Œuvres complètes*, vol. 1, Paris.
- Buckland, 1998 : Michael Buckland, « What is a "Digital Document" ? », *Document numérique*, 2 (2), p. 211-230.
- Febvre et Martin, 1999 : Lucien Febvre et Henri-Jean Martin, *L'Apparition du livre*, 3e édition, Paris.
- Goldfarb, 1991 : Charles F. Goldfarb, *The SGML Handbook*, New York.
- Haigh, 2006 : Thomas Haigh, « Remembering the Office of the Future : The Origins of Word Processing and Office Automation », *IEEE Annals of the History of Computing* , 28 (4), p. 6-31.
- Knuth, 1999 : Donald E. Knuth, *Digital Typography*, Stanford.
- Levy, 2001 : David M. Levy, *Scrolling Forward : Making Sense of Documents in the Digital Age*, New York.
- Martin, 1988 : Henri-Jean Martin, *Histoire et pouvoirs de l'écrit* , Paris.
- Marshall, 2003 : Catherine Marshall, « Reading and Interactivity in the Digital Library : Creating an Experience that Transcends Paper », *In Proc. of the CLIR/Kanazawa Institute of Technology Roundtable*, Kanazawa, Japon, p. 1-20.
- Ranganathan, 1963 : Shiyali Ramamrita Ranganathan, *Documentation and its Facets*, Londres.
- Rubinstein, 1988 : Richard Rubinstein, *Digital Typography : An Introduction to Type and Composition for Computer Design*, Reading.
- Sellen et Harper, 2002 : Abigail J. Sellen et Richard H. R. Harper, *The Myth of the Paperless Office*, Cambridge (Mass.).

Nos partenaires

Le projet *Savoirs* est soutenu par plusieurs institutions qui lui apportent des financements, des expertises techniques et des compétences professionnelles dans les domaines de l'édition, du développement informatique, de la bibliothéconomie et des sciences de la documentation. Ces partenaires contribuent à la réflexion stratégique sur l'évolution du projet et à sa construction. Merci à eux !

- CONCEPTION : ÉQUIPE *SAVOIRS*, PÔLE NUMÉRIQUE RECHERCHE ET PLATEFORME GÉOMATIQUE (EHESS).
- DÉVELOPPEMENT : DAMIEN RISTERUCCI, IMAGILE, MY SCIENCE WORK.
- DESIGN : WAHID MENDIL.