Le SGML et son intérêt pour la gestion des documents juridiques

Daniel Poulin Professeur, CRDP, Faculté de droit, Université de Montréal

---

Introduction

Au cours des quinze dernières années, le monde de la documentation juridique a vécu d'importantes transformations. Essentiellement, les juristes sont passés du papier à l'électronique. Aujourd'hui, dans la plupart des cas, nous disposons de deux supports matériels complémentaires pour conserver et diffuser la documentation juridique. Si le papier demeure irremplaçable pour la lecture studieuse, il cède peu à peu sa place à son vis à vis électronique pour toutes les autres formes d'utilisation de l'information. En effet, l'arrivée des médias électroniques a transformé notre relation à l'information, de nouvelles attentes sont nées. Nous voulons désormais des systèmes qui permettent le repérage et simplifient la réutilisation de l'information. Nous prenons de plus en plus pour acquis les bénéfices du document électronique. Cependant, cette transition du papier à l'électronique n'est pas terminée. Au contraire, elle se poursuit et, pourrions-nous ajouter, le meilleur est encore à venir. Les prochains pas devraient augmenter considérablement les bénéfices des médias électroniques.

Un élément clé des changements à venir réside dans l'adoption de nouveaux encodages - de nouveaux formats - pour nos documents. Ceux-ci devraient nous permettre d'échanger plus commodément l'information et de la publier à meilleur compte. L'encodage cherché devrait également mieux assurer la pérennité des documents, c'est-à-dire, qu'il devrait allonger leur durée de vie. Un format de ce type est disponible, il s'agit du SGML (Standard Generalized Markup Language). Certains estiment depuis longtemps son adoption inévitable. Toutefois, l'intérêt d'y recourir n'est devenu parfaitement clair qu'avec le développement récent d'Internet.

En deux ans à peine Internet a bouleversé le monde de l'informatique, celui des médias et de l'édition. Le souffle principal de la tornade Internet vient du World Wide Web -- le Web, cet immense réseau de serveurs d'information qui couvre aujourd'hui la planète. Le succès du Web tient dans une bonne mesure à l'espace informationnel unifié, indépendant des logiciels et des ordinateurs, qu'il a permis de créer. Ce faisant, le développement du Web a fourni une preuve d'une clarté exceptionnelle de l'intérêt des approches normalisées pour la publication et la transmission des documents. En effet, bien qu'Internet ait couvert la majeure partie du monde depuis des années, ce n'est qu'avec l'arrivée du Web qu'il a littéralement explosé pour devenir un phénomène technologique, culturel et social. La réussite du Web tient à trois éléments : une méthode normalisée d'adressage (les URLs), un protocole de communication (le HTTP) et, c'est ce qui nous intéressera ici, à une convention de balisage, le HTML (Hypertext Markup Language) qui n'est en fait qu'une utilisation particulière du SGML. Le HTML a établi mille fois mieux que toutes les démonstrations théoriques l'utilité de préparer ou baliser l'information de façon normalisée.

Le balisage ou le marquage des textes

Le *balisage* ou le *marquage* des textes n'est pas une idée nouvelle, en fait, les espaces blancs séparant les mots d'une phrase peuvent être vus comme une forme de marquage. La ponctuation aussi est un marquage, tout comme le sont les éléments de présentation, la séparation des paragraphes et la pagination. Lors du passage à l'électronique, au plan du marquage des textes, nous avons reproduit les habitudes acquises à l'époque du papier, de la dactylo et de l'écriture manuscrite. Pour l'essentiel, les traitements de textes ont été utilisés pour reproduire le marquage traditionnel coloré, peut-être, de quelques nouvelles marques comme la variation de la taille des polices de caractères.

Le *marquage logique*, ou *descriptif*, à la base du SGML reflète un autre point de vue. Pour les partisans du marquage logique, il importe surtout que le marquage nous renseigne directement sur la structure des documents, ses parties logiques, et non sur l'apparence qui devrait révéler plus ou moins clairement cette structure. Dans un cas, le marquage indique que certains mots doivent apparaître en caractères gras; dans l'autre, il informe que certains mots sont un titre; quitte, bien sûr, à ce que l'on informe par ailleurs le logiciel que nous souhaitons voir les titres s'afficher en caractères gras.

Avec le *marquage logique* ou *descriptif*, le document comporte deux types d'informations, d'une part, le *contenu* lui-même, c'est-à-dire, les caractères, les mots, les illustrations, et, d'autre part, l'information sur la structure, c'est-à-dire, les *balises*. Cette information descriptive servira par exemple à identifier les chapitres, les titres, les références. De cette façon, de nouvelles utilisations du texte sont possibles. De meilleurs systèmes de repérage peuvent être conçus. En particulier, il devient possible de concevoir des systèmes exploitant la structure des documents où l'usager peut, par exemple, spécifier dans ses requêtes que les termes cherchés doivent se trouver dans les titres. Le marquage logique facilite aussi le passage d'un média à l'autre. Le document n'étant pas encombré de caractéristiques typographiques liées à une seules de ses utilisations possibles, le document balisé logiquement peut être publié sur papier, sur un serveur Web ou sur un CD-ROM. À chaque fois, il suffira d'indiquer comment chaque type d'éléments devra apparaître. D'autres avantages encore peuvent être réalisés et les prochains paragraphes tenteront de les mettre en lumière. Ces avantages du balisage logique, pour être maximisés, exigent que le balisage soit standardisé, d'où l'intérêt du SGML.

Le SGML

Le SGML (*Standard Generalized Markup Langage*) a été élaboré au début des années quatre-vingt. Il s'agit d'un langage informatique formel permettant la description des documents. C'est aussi un langage normalisé, en ce sens qu'il a été adopté par l'ISO (*International Standards Organization*). Notons que SGML n'appartient à aucun manufacturier, et pour cette raison, il fournit un moyen d'encoder l'information sur support électronique sans qu'elle ne dépende d'aucun logiciel spécifique. Ainsi, les documents SGML peuvent être échangés à travers le monde, d'un ordinateur à l'autre, d'un logiciel supportant le SGML à un autre. Par ailleurs, le marquage des documents est formel et il peut être validé. Il devient donc possible de vérifier qu'un document est *correct* ou, en d'autres termes, qu'il est construit comme il devrait l'être. Pour ces raisons,

plusieurs estiment qu'il s'agit de la langue commune du monde documentaire de demain. De façon plus pratique, quatre bénéfices principaux sont associés au SGML:

* la protection de l'investissement (l'organisation ne se trouvera pas coincée dans la technologie d'un vendeur particulier);

* la réutilisation des données (les marques ou balises étant associées aux éléments structurels du document pour expliciter leur rôle, l'information peut être plus facilement repérée et réutilisée);

* la sécurité et la pérennité de l'information (l'utilisation d'un format stable, n'appartenant à aucun manufacturier facilite l'archivage);

* l'interopérabilité (SGML est la "lingua franca" des systèmes ouverts d'information comme Internet).

Dans le contexte actuel, celui où les inforoutes s'étendent rapidement, où les exigences d'accessibilité de l'information sont croissantes, mais où, au même moment, les ressources financières des États sont décroissantes, il n'est pas surprenant de constater que dans bon nombres de pays les partisans de l'utilisation de SGML se multiplient. D'ores et déjà, en Europe et aux États-Unis, le SGML est utilisé par plusieurs organismes gouvernementaux et de nombreuses législatures, dont la CEE, le HMSO de Grande-Bretagne, le DOD, le DOE, le GPO, l'IRS, la SEC à Washington[1]. Il en va de même dans un nombre grandissant d'États américains, dont ceux du Texas, de l'Oklahoma, du Rhode-Island, du Wisconsin, du Maryland et de l'Iowa. Plus près de nous, le SGML vient d'être retenu par le ministère de la Justice du Québec et la société SOQUIJ pour la conception du niveau système d'information de la magistrature québécoise.

Le document SGML

Tout document comporte implicitement trois types d'information, des données, une structure et des directives de formatage[2]. Les données constituent la partie la plus connue du document, il n'est pas nécessaire de s'y attarder ici. La *structure* exprime les relations entre les divers éléments du document. Quant aux *formats*, ce sont eux qui conditionnent l'apparence du document : polices de caractères, marges, espacements, mise en page. Le SGML reconnaît, exprime et exploite ces distinctions. Il permet de préserver le contenu et la structure sans spécifier de format, reportant l'introduction de cet élément au moment où sera choisi le support de livraison et déterminés les besoins d'une clientèle spécifique.

Le document SGML comporte lui aussi trois parties : la déclaration SGML, la DTD (Définition de Type de Document) et le corps du texte. La première, la *déclaration SGML* , permet entre autres de préciser le jeu de caractères et le mode de marquage du document. Elle est le plus souvent omise parce que sa version de défaut est invoquée automatiquement par les logiciels SGML.

Le deuxième élément du document SGML est la DTD ou définition de type de document qui définit la structure logique du document. Cette déclaration exprimée dans un langage formel. Elle figure en tête du document ou encore elle est, elle aussi, simplement évoquée pour que le logiciel lecteur puisse la retrouver. En fait, la *DTD* est une description abstraite d'une classe de documents. Pour la construire, il nous faut identifier les éléments qui peuvent composer le document. Une fois, les éléments acceptables identifiés, la DTD précise leurs combinaisons autorisées. La granularité de la DTD peut être plus ou mois fine, en ce sens

que le niveau de détails peut varier d'une DTD à l'autre selon les besoins considérés. Plusieurs DTD sont donc possibles pour une même classe de documents. Quoiqu'il en soit de sa granularité, le rôle de la DTD sera double. D'une part, elle contraint plus ou moins strictement les formes acceptables des documents tandis que, d'autre part, elle permet d'identifier la structure du document pour l'exploiter. Une DTD particulière pourra, par exemple, spécifier les rubriques d'un document devant obligatoirement être remplies.

Le *corps du texte* véhicule l'information constituant la raison d'être du document. Son contenu ressemble beaucoup aux documents que nous connaissons déjà à cette différence près que des balises y ont été insérées pour délimiter la structure logique du texte. Ces balises portent les noms retenus par le concepteur de la DTD. Généralement, ces noms figurent entre crochets et les balises viennent par paires. Pour des fins d'illustration, le prochain paragraphe a été balisé comme un élément de type paragraphe.

<pre><paragraphe></pre>On peut associer des attributs aux balises pour qualifier le texte qu'elles délimitent. Les valeurs que peuvent prendre ces attributs auront été définies lors de la préparation de la DTD. Un attribut de langue pourra par exemple spécifier que la langue d'un paragraphe ou d'une citation n'est pas celle du document -- la valeur par défaut de l'attribut -- mais une langue étrangère, l'italien ou l'allemand. Par la suite, on pourra exploiter cet attribut pour repérer les citations en langues étrangères dans un document ou un corpus entier.<pre></paragraphe></pre>

Comme un texte ainsi enrichi peut devenir difficile à lire, les logiciels de lecture dissimulent généralement ces balises tout en s'en servant pour rehausser la présentation du document. En fait, la norme SGML peut être rendue invisible à l'utilisateur dans le contexte d'une application concrète.

Cette combinaison de marques et de contenus n'est pas nouvelle. Ceux qui utilisaient le WordPerfect il y a quelques années se rappellent que l'on y inscrivait très explicitement des marques d'italique, de souligné et autres. Ces marques ressemblent à celles que l'on insère avec le SGML. Cependant, à la différence des marques insérées par WordPerfect celles associées au SGML sont destinées à expliciter la structure des documents.

Même balisés en SGML, les documents doivent à un certain point de leur existence trouver une apparence typographique acceptable. Pour y parvenir, un fichier fournissant les caractéristiques typographiques associés aux divers éléments structurels ou logiques doit être préparé. Différentes méthodes permettent de définir l'apparence d'un document SGML sur un support particulier. Dans l'avenir, une approche normalisée sera disponible pour associer le document SGML à sa publication sur un média particulier.

Les avantages de cette séparation des contenu, structure et apparence sont considérables. Tout d'abord, si on souhaite changer l'apparence de toute une série de documents, il n'y qu'à modifier un seul fichier celui où est précisé l'apparence des éléments logiques. Il n'est pas nécessaire de formater chacun des fichiers contenant les documents. Cette possibilité est particulièrement avantageuse lorsque l'on envisage de diffuser le même texte au moyen de plusieurs médias, papier, CD-ROM et Internet.

L'intérêt du SGML pour la documentation juridique

Les normes documentaire comme le SGML peuvent avoir un impact considérable sur la vie de la communauté juridique. La diffusion et les échanges d'information jouent en effet un rôle central en droit.

Ces nouvelles normes recèlent le potentiel de transformer tant la façon de concevoir et d'utiliser la documentation juridique que la façon de conduire des échanges entre juristes.

L'adoption par les éditeurs, par les tribunaux, voire par les législateurs, du SGML entraînera l'apparition de plusieurs nouveaux produits informationnels et ce sur divers médias. En effet, alors qu'avec les approches traditionnelles chaque publication dérivée d'une collection de documents bruts entraîne des coûts importants; au contraire, le recours au marquage logique facilite, lui, la réutilisation de documents ou d'éléments de ces documents dans de nouveaux contextes d'édition. La réduction du coût de développement de nouveaux produits informationnels spécialisée ne peut qu'être bénéfique. En particulier, il sera moins coûteux et plus simple de produire des publications électroniques dotées de mécanismes sophistiqués repérage. Enfin, la conservation du patrimoine juridique sur support électronique sera mieux assurée.

La popularisation des normes documentaires peut aussi rendre des services considérables aux juristes lors de leurs échanges de documents. Une des difficultés de l'échange de documents juridiques tient au fait que l'on souhaite combiner les avantages d'une structure rigoureuse et contraignante, que le message d'EDI formulé selon les normes internationales sert si bien, à ceux du texte libre, le véhicule traditionnel des échanges en droit. Le SGML et les normes qui y sont associées offrent précisément une solution propre à concilier ces exigences contradictoires. En effet, la forme du document SGML peut être pratiquement aussi libre qu'un texte en langage "naturel" ou, tout au contraire, tout aussi restrictif qu'un message d'EDI[3] selon les préoccupations qui président à la conception de la DTD. Il devient donc possible de concevoir des formes d'échange combinant des éléments rigoureusement déterminés et d'autres beaucoup plus libres. Il semble donc que le SGML peut devenir la pierre d'assise des normes d'échange en droit. Cela n'est pas surprenant, car l'échange d'information et de documentation est à l'origine même du SGML[4].

Conclusion

Certains comparent le changement résultant par les nouvelles technologies de l'information à la révolution déclenchée par les travaux de Gutenberg il y a plus de cinq siècles. Sommes-nous à la veille d'une révolution des modes de circulation de l'information dans le monde juridique? Certains le pensent. Sur ce point, une anecdote récemment citée par Nicholas Negroponte vaut la peine d'être reproduite[5] . Negroponte reprend cette devinette où un ouvrier accepte un emploi à un sous par jour la première journée à la condition qu'on l'embauche pour un mois et ... que son salaire soit doublé à chaque jour. L'employeur naïf saute sur l'aubaine loin de deviner qu'à la fin du mois il aura dû lui verser plus de 10 millions de dollars. C'est alors que Negroponte nous fait observer que les derniers jours de l'emploi sont d'une extrême importance. Un contrat semblable d'une durée de trois semaines n'aurait coûté que vingt mille dollars à l'employeur. Il en va de même selon lui des transformations dont nous serons témoins dans le monde des médias et, pourrions-nous ajouter, plus spécifiquement, dans celui de la documentation juridique. Nous en sommes au point où les progrès technologiques des vingt dernières années commencent à faire sentir leur effet cumulatif.

Une dernière question doit être abordée. Le SGML appartient-il à une nouvelle mode appelée a être à nouveau remplacée dans peu de temps? La science de la prédiction est une discipline fort incertaine. Néanmoins, sur ce point, l'exemple récent du développement du Web peut nous être utile pour évaluer le potentiel du SGML. Le SGML a été conçu au début des années quatre-vingt et personne n'entrevoyait alors l'émergence du Web. Pourtant, une dizaine d'années plus tard, le SGML est au centre de cet immense espace informationnel. Il ne s'agit pas là d'un hasard ou d'une chance. La mise en oeuvre universelle du HTML, qui n'est qu'une DTD

particulière, s'explique par la nature très ouverte du SGML. Le SGML est en effet un langage extrêmement général capable d'être adapté à de nombreuses situations nouvelles. Il a été conçu pour cela. Bien sûr, nul ne peut prédire l'avenir. Tout au plus, pouvons-nous choisir les solutions conçues pour s'adapter au changement, et c'est cela que nous offre le SGML.

Cybernews Volume 2, numéro 3 (automne 1996)

---

1 La CEE (Communauté économique européenne), le HMSO (Her Majesty Stationnary Office) de Grande-Bretagne et, aux États-Unis, le DOD (Department of Defence), le DOE (Department of Energy), le GPO (General Printing Office), l'IRS (Internal Revenue Service), la SEC (Securities & Exchange Commission).

2 Bien que traditionnellement les documents juridiques se soient limités à l'utilisation du texte, on peut penser que dans l'avenir les *données* pourront être de plusieurs types: texte, tableaux, graphiques, images et même, pourquoi pas, objets multimédias comportant son et images animées.

3 On trouvera un exemple de formalisation d'un message d'EDI en SGML dans Eric Van Herwijnen, "SGML Pratique", International Thompson Publishing, Paris, 1995, 330 p.

4 Il faut noter que la première dissémination importante du SGML a fait suite à un projet d'échange d'information mis sur pied par le Pentagone dans le cadre du projet nommé JCALS (Joint Continuous Acquisition and Lifecycle Support). Il s'agissait alors de permettre à l'information associée au matériel militaire de circuler en format SGML à l'intérieur des forces armées et aussi, bien sûr, entre elles et leurs fournisseurs. Ce projet a fait école. Par la suite, des secteurs industriels entiers ont adopté le SGML à des fins similaires. Les compagnies aériennes et l'industrie automobile ont élaboré des DTD s'appliquant tant à leur production de documents et de manuels d'entretien, qu'à leurs appels d'offre, et à leurs soumissions

5 L'homme numérique, Robert Laffont, Paris, 1995

 ISSN 1201-7302
