

Les normes de publication et consultation des traductions

Introduction

Ces dernières années, le problème de *la pérennisation de la communication des documents mono et multilingues* s'est posé de façon accrue.

Les évolutions technologiques sont si rapides qu'un titre édité sur un support CédéRom pour une certaine plate-forme ne pourra probablement plus être lu dans les environnements qui sortiront dans cinq ans. Une des solutions pourrait être la migration fréquente des applications ou des documents multilingues multimédia vers les plates-formes plus récentes, à condition d'avoir effectué le développement dans un standard pérenne.

Un autre problème auquel se voient confrontés les traducteurs au sein d'organisations industrielles ou gouvernementales est celui lié à l'accroissement du volume d'informations traitées, stockées et partagées. Sur la chaîne de production de cette information, coexistent plusieurs intervenants, et l'information créée est mise à disposition et restituée au plus grand nombre de lecteurs potentiels. Le problème de ces applications fonctionnant en réseaux est celui de la réalisation d'applications indépendantes des plates-formes d'exécution.

Là encore la solution passe par l'adoption de règles de développement et de codage des applications et des documents multimédias monolingues ou multilingues qui soient des standards ou des normes.

L'objectif de ce premier point est de rappeler les normes ou standards de codage en faisant ressortir leur champ d'application, leur niveau d'introduction dans le marché visé et leur positionnement relatif.

2. Types de normes

Il faut distinguer entre les *normes de codage/norme de production de documents* et les *normes d'échange de données, de documents ou d'applications multimédia*.

Ces normes standards peuvent intervenir à plusieurs niveaux : interface utilisateur, format des données à échanger, bases de données (stockage, accès, requête, etc.), protocole réseaux. Nous intéresserons au codage numérique de la structure du document avec les normes SGML (norme ISO), HTML (standard IETF) et XML, puis à PDF, norme pour les formats des données à échanger. Chaque standard sera présenté avec ses caractéristiques principales, un bref historique, le segment de marché visé et pénétration et ses principales techniques de base.

3. Normes de production des documents

Le document monolingue ou multilingue est avant tout sous forme électronique, sa présentation papier n'étant qu'un des formats de restitution possible. Pour un document électronique, il n'existe pas de séparation entre le contenu du document, sa structure et sa représentation.

La structure logique d'un document électronique peut être décrite à l'aide d'un jeu de balises qui permet de distinguer les différentes parties d'un document (titres, sous-titres, etc.). Les normes SGML, HTML et XML permettent le balisage logique de l'information contenue dans le document. Ainsi, on obtient des documents dits structurés.

3.1. Standard Generalized Markup Language (SGML)

3.1.1. SGML à l'origine

SGML fut créé en 1978 par le American National Standards Institute (ANSI) et la Graphic Communications Association (GCA) et devint une norme en 1983. Il permet de représenter la logique du contenu d'un document, avec des inclusions possibles d'objets graphiques. C'est un langage qui s'adresse surtout au domaine de l'édition, des grandes entreprises devant gérer, produire et diffuser une grande quantité de documents monolingues et multilingues comme celles du secteur de l'aéronautique ou de l'automobile. Il permet de structurer les textes et de lier des bases de données aux documents. Un document SGML ne comprend ni d'information de style ni de mise en page.

Il comporte une séparation entre le balisage du contenu pour représenter sa structure et l'information sur son apparence.

3.1.2. Présence sur le marché de SGML

C'est un langage qui s'est imposé dans le domaine de l'édition, dans les grandes entreprises (par exemple dans le domaine de l'aéronautique), institutions gouvernementales et européennes (publications officielles de l'Union Européenne). SGML permet de garantir la pérennité de l'information en utilisant un langage indépendant des outils d'affichage et de formatage. De nombreux outils de conversion ou de création existent.

De nombreux organismes supportent le développement de SGML comme le SGML User's Group ou SGML Open.

3.2.3. Principes techniques

SGML représente la structure logique des documents électroniques plutôt que la forme de leur présentation. Les documents générés ont une structure de données logiques, c'est-à-dire que l'on définit la structure logique du document (titre, sous-titre, corps de texte, etc.), mais pas sa forme (titre en gras ou en italique par exemple),

Un document SGML peut être défini d'une part par sa structure et d'autre part par son contenu. A chaque document SGML on associe une DTD (Document Type Definition), une grammaire, interne ou externe qui décrit la structure du document. Cela peut ressembler à la façon dont on décrit une base de données. Chaque DTD regroupe un jeu de balises, les principes de balisage choisis ainsi que les règles syntaxiques associées. Le nombre et la variété des DTD sont infinis. Une DTD fournit le cadre pour organiser les différents éléments d'un document comme les chapitres, sections et entêtes. On distingue les DTD conceptuelles, où les éléments sont des informations de métier, des DTD éditoriales, où les éléments sont des informations de texte, chapitre, section, tableau, etc.

Comme SGML ne définit pas les règles de style, cette fonction est remplie par une autre norme ISO, la DSSSL (Document Style Semantics and Specification Language).

SGML peut utiliser des objets non SGML qui sont cependant définis comme tels. Dans ce cas, il existe des limitations, en particulier en ce qui concerne le traitement d'éléments multimédias et la liaison de documents différents.

SGML ne peut pas imposer la façon dont sera interprété le jeu de balises. L'interprétation de balise dépend de l'application de traitement de l'interprétation.

4. Hypertext Markup Language (HTML)

4.1. HTML en général

Dérivé du standard SGML, HTML fut développé au Cern par Tim Berners-Lee. Il s'impose comme langage pour le Web en 1990. Il permet la création de documents hypertextes portables d'une plate-forme à une autre. Afin de publier de l'information sur le Web, il était nécessaire d'avoir un langage universellement compréhensible par tous les ordinateurs.

HTML est un langage de balisage très simple qui s'apprend rapidement. Les liens sont également simples à mettre en place, une simple balise <A...> balisant à la fois la source et la cible du lien. Enfin, HTML est facilement portable.

Cette simplicité a cependant des limites. Un balisage d'éléments sémantiques n'est pas possible ce qui limite les performances des moteurs de recherche.

L'ISO/IEC soutient son développement.

4.2. Principes techniques

HTML est doté de fonctions multimédias, de langages de scripts, de feuilles de style, de cadres (frames), de tableaux et de formulaires ainsi que de fonctionnalités pour les handicapés, tout en gardant un format simple par refus d'adopter la complexité des applications SGML classiques.

Il est possible d'inclure des objets multimédias et la présentation des pages est simplifiée grâce aux feuilles de style. Une feuille de style permet de définir un format standard pour l'affichage des pages d'un document : police de caractères, alignement, couleur, etc.

La feuille de style peut être spécifiée pour un élément ponctuel ou pour un groupe d'éléments. Celle-ci peut être placée à l'intérieur d'un document HTML ou dans une feuille de style séparée.

Le système associe une feuille de style à un document indépendant du langage de la feuille de style.

Il est possible d'écrire dans toutes les langues (écrire de gauche à droite et de droite à gauche) grâce à l'adoption du standard ISO/IEC10646 qui est la table standard pour HTML.

Les feuilles de style permettent de présenter les pages en fonction des préférences des lecteurs / utilisateurs finals.

Le document peut être interprété par diverses plates-formes. Le lecteur de langue braille ou synthèse vocale.

Comme pour SGML, il est recommandé de séparer la structure de la présentation.

5. Extensible Markup Language (XML)

5.1. XML en général

XML ou langage extensible de balisage de document contenant de l'information structurée. Il est basé sur SGML et Hytime **faut-il remettre le footnote?** pour le Web. Il sera le successeur de HTML.

Ce langage de balisage présente l'information (texte, graphique, image fixe, etc.) en intégrant des balises qui définissent sa structure.

Il se positionne entre HTML et SGML. Il allie à la fois les points forts de SGML et la simplicité de HTML.

XML est la forme limitée de SGML et est plus facile à mettre en œuvre au niveau du Web. Il étend les capacités sémantiques de HTML.

XML interagit avec SGML et HTML.

Comme pour SGML, il est possible de séparer le contenu et la structure du document dans XML.

Aucun élément de mise en forme ne se trouve dans le code de XML. Comme pour SGML, l'information contenue dans un document est considérée comme une donnée et peut donc être gérée dans une base de données.

Enfin, tout comme pour SGML, XML permet l'écriture de DTD et permettant ainsi une personnalisation du balisage pour tout type de document.

5.2. Bref historique de XML

XML vit le jour grâce au travail effectué au sein du consortium W3C en 1996 dont l'objectif était de développer un langage pour les applications Web palliant la simplicité de HTML et la complexité de SGML. La première recommandation XML 1.0 parut en 1998.

5.3. Présence sur le marché

Il est encore tôt pour établir un bilan puisque ce langage est tout neuf.

XML pourrait se concentrer sur des questions de bibliothéconomie. Le problème essentiel du Web est la recherche de l'information au milieu de milliards de pages. Toutes les pages ne sont pas identifiables sans l'aide de moyens d'indexation et de documentation automatiques. XML apporte un début de solution au catalogue et aux informations. En effet, en structurant les données, XML permet aux agents intelligents de rechercher efficacement les données nécessaires au catalogage qui sont contenues dans les pages. XML s'étendra à d'autres applications puisqu'il est soutenu par Microsoft.

XML comprend trois parties qui sont les suivantes :

- La DTD d'un document XML peut comporter ou non une DTD. S'il comporte une DTD, alors le document est dit valide, et il doit la respecter. Si le document ne respecte pas une DTD, alors il doit au moins respecter un formalisme plus précis que HTML. Par exemple, les éléments peuvent être vides et il y a systématiquement des balises d'ouverture et de fermeture.
- Le XSL (eXtensible Stylesheet Language) est un langage qui définit les styles associés à un document XML. Avec XSL, un document possède des attributs de mise en forme : fonte, couleur, etc. Il permet de formater les documents en tenant compte d'éléments d'héritage, d'intégrer des textes et des graphiques, de réutiliser des macros, etc : Plusieurs vues du même document peuvent être présentées à différents utilisateurs.
- Le XLL (eXtensible Linking Language). Les liens hypertextes en XML sont décrits en utilisant le langage XLL. XLL est basé sur la norme Hytime ¹. **Doit-on remettre le footnote?** Il permet des liens internes au document ou externes vers d'autres documents. XLL est composé de deux parties : Xlink et Xpointer. Xlink spécifie les liens qui peuvent être insérés dans XML, et Xpointer la façon d'adresser et d'identifier les objets liés dans un document XML.

Les liens XML ne sont pas de simples liens hypertextes. Ceux-ci prennent en compte les éléments suivants.

- Liens multidirectionnels qui permettent de revenir au point de départ d'un lien, ce qui n'était pas possible avec HTML
- Liens avec destinations multiples à partir d'un même point de départ
- Inclusion dans le corps d'un document d'une partie de document lié qui pourra être ainsi mise à jour de façon transparente.

Des études ont démarré pour intégrer des éléments VRML à un document XML. Une DTD spécifique est également à l'étude.

Les graphiques dans XML sont rattachés au document XML par des liens. Ce sont des liens Xlink et Xpointer qui associent les images graphiques. L'activation des liens permet d'afficher à l'aide d'un clic l'image au chargement ou encore dans une fenêtre.

XML ne restreint pas l'utilisation d'un format graphique. Ainsi, il est possible d'utiliser des images GIF, JPEG, TIFF, CGM, etc.

Après avoir vu les normes relatives à la production de documents, passons maintenant aux normes qui permettent de diffuser les documents via le Web pour consultation.

6. Diffusion de documents pour consultation

Par consultation, nous entendons : lecture seule du document, sans y apporter directement de modification. C'est le cas des documents mis à disposition sur un site Web ou sur un site ftp public, ou envoyé pour information à une communauté d'utilisateurs (exemples: compte-rendus). Pour diffuser un document, nous présenterons les deux formats les plus utiles pour les traducteurs PDF, MIME.

6.1. Portable Document Format (PDF)

6.1. Définition

Portable Document Format (PDF) est un format d'échange de documents développé par la firme Adobe. Pour lire un document PDF, vous avez besoin du logiciel AcrobatReader. Celui-ci est téléchargeable gratuitement sur le site de Adobe. Cependant, le logiciel qui permet de rédiger des documents PDF est payant.

Un document PDF est un document en plate-forme croisée autonome. Il s'agit simplement d'un fichier qui se présente de la même façon sur l'écran et en imprimé, indépendamment du type de PC ou d'imprimante utilisés et indépendamment du logiciel qui a été conçu à l'origine pour le créer.

Bien qu'ils contiennent la mise en forme complète du document original, y compris les polices de caractères et les images, les fichiers PDF sont très condensés, ce qui permet de télécharger efficacement des informations complètes.

6.2. Avantages de PDF par rapport à PostScript

- format **indépendant** de la plate-forme et de l'application créatrice
- tout document peut être imprimé sur toute imprimante, PostScript ou non, en utilisant le mécanisme d'impression standard de la plate-forme utilisée

- format beaucoup plus compact que le même document en format PostScript : des rapports de 1 à 10 sont fréquents, dépendants du contenu du document et des taux de compression utilisés
- technologie de substitution intelligente de fontes, liens hypertexte, table des matières (*bookmarks*), possibilité d'ajout de notes en superposition, de couper/coller avec d'autres applications, etc.
- documents lisibles à l'écran avec possibilité de zoom variable
- adaptation automatique au format de sortie avec réduction si nécessaire (pas de problème de bac ou de format de papier non disponibles).

6.2. Multi-purpose Internet Mail Extensions (MIME)

6.2.1. Définition

MIME, l'**extention** "Multi-purpose Internet Mail Extensions", est une spécification décrivant les formats de messages multimédias sur l'Internet. MIME est disponible et utilisable gratuitement.

Avec MIME, il est possible d'envoyer et de consulter des messages ou des documents comprenant :

- des jeux de caractères autres que l'ASCII
- du texte enrichi
- des images
- des sons
- des séquences vidéo
- des condensés de courriers
- des "pointeurs" de fichiers (URL) .

MIME est doté de deux formats de codage QP et base 64 qui permettent d'utiliser des alphabets étendus sur 8 bits (et même plus), ouvrant ainsi les possibilités d'envoyer des messages par courrier électronique dans un grand nombre de langues.

MIME ne restreint pas son champ d'action aux deux formats de codage (QP et base64) qu'il introduit. Un logiciel supportant MIME est capable d'interpréter des messages dont le contenu n'a pas été codé, comme ceux véhiculés en "7 bit", "binary" ou "8 bit". Parmi ceux ci, les deux derniers peuvent être la source de problèmes s'ils traversent des dispositifs du réseau qui tronquent les messages au-delà du 8ème bit.

Conclusion

En concluant sur les normes de publication, nous pouvons dire que la production de documents multilingues destinés à être balisés en SGML ou XML doit faire l'objet de recommandations dont plusieurs doivent aller vers les comités de relecture des traductions et

des formateurs des réviseurs. Il faudrait donc définir ou renforcer, pour chaque type de document à traduire, un modèle normalisé quant à la présentation des textes et l'appliquer de façon rigoureuse. Cet aspect d'importance capitale permettrait d'élaborer un document du type « Guide ou protocole de rédaction » et encouragerait les auteurs ou les traducteurs à utiliser une feuille de style calquée sur la DTD, laquelle serait rendue disponible par le commanditaire sur son site Web ou encore directement envoyée aux auteurs ou traducteurs par courrier électronique.

Comme beaucoup de modifications, de corrections, d'ajouts se font au cours de la traduction et prennent beaucoup de temps, à long terme, il faudra envisager dès le début de la chaîne de traitement que le traducteur travaille le document à traduire dans un éditeur SGML ou XML et produise ainsi une version SGML ou XML de la traduction.

Les avantages d'un éditeur SGML/XML sont nombreux : validation de la syntaxe SGML/XML, listes d'autorités et alias pour des valeurs attributs et de contenus d'éléments, environnement personnalisé de création, direction dans la structuration du texte en indiquant quels éléments sont permis à quels emplacements de la traduction, etc. Ceci permet au traducteur de gagner du temps et d'obtenir des résultats plus homogènes (tant inter qu'intra documentaires). Le produit obtenu est directement du SGML ou du XML et n'a pas besoin de conversions supplémentaires. Le travail de diffusion à investir en aval de la production de la traduction est considérablement réduit.

PDF permet de consulter aisément une grande quantité de documents. Il n'est cependant pas possible de corriger le document directement car celui-ci est en mode "lecture".

Références

A) Références concernant directement les normes

Hypertext Format language, norme du World Wide Web Consortium. Voir : <http://www.w3.org/MarkUp>

Pour XML , Voir : <http://www.w3.org/TR/1998/REC-xml-199980210>

Pour les Cascading Style Sheet, Voir : <http://www.w3.org/Style/css>

Pour obtenir une liste exhaustive concernant les convertisseurs en SGML, consulter la rubrique « Conversion Program » dans Survey of software for structured text, par Eila Kuikka et Erja Nikunen. Voir : <http://www.uku.fi/~kuikka/systems.html>

Un programme de conversion en SGML nommé OmniMark, développé par Rick Geimer, est disponible gratuitement sur le Web. Voir : <http://www.xmeta.com/omlette>

B) Références bibliographiques

Kasdorf, Bill (1998) : *SGML and PDF – Why ye need both*. In. Journal of electronic publishing, Vol. 3, No 4.
Disponible sur Internet à l'URL. <http://www.press.umich.edu/jep/03-04/kasdorf.html>

Lieb, Thom (1999) : *HTML, PDF ans TXT : the format wars*. In : Journal of electronic publishing : Vol . 5, No 1.
Disponible sur Internet à l'URL. <http://www.press.umich.edu/jep/05-01/lieb501.html>

Marcoux, Yves (1994). *Les formats normalisés de documents électroniques*. ICO Québec.
Vol : 6, No1-2-, p. 56-65.
Disponible sur Internet à l'URL. <http://tornado.ere.umontreal.ca/~marcoux/grds/ico94.htm>