SGML, une norme pour les autres, mais pas pour moi ?

Françoise ROSSION

Qui n'a pas, un jour, banni son traitement de texte ou son micro-ordinateur parce que le fichier qu'il souhaitait manipuler était devenu illisible ou avait perdu certaines informations structurelles importantes ?

Depuis que l'informatique existe, les problèmes de compatibilité sont le lot quotidien des informaticiens amateurs que nous sommes. Chaque jour, de nouveaux ordinateurs, des logiciels inédits apparaissent sur le marché informatique afin de répondre aux souhaits d'utilisateurs de plus en plus exigeants. Ainsi, se développe un vaste parc de machines et de programmes incapables de « dialoguer » entre eux.

La saisie de textes sur support informatique n'est qu'un exemple de ce que le terme « incompatibilité » signifie : encoder son article sous *WordPerfect* et l'imprimer tel quel le lendemain à partir de *MS-Word* relève souvent du défi.

Les éditeurs américains se sont, les premiers, penchés sur ce problème; ils furent bientôt suivis par la communauté internationale : la norme SGML (*Standard Generalized Language Markup*) est le résultat de ces efforts conjugués.

✉ Université Libre de Bruxelles; Section des Sciences de l'Information et de la Documentation; INFODOC; C.P. 142; 50, Avenue F.D. Roosevelt; B–1050 Bruxelles (Belgique).
Fax : + 32 2 6503921

MOTS-CLÉS : SGML, normalisation, document, application, informatique, base de données, description structurelle, norme internationale.

Histoire d'une norme

SGML est un standard reconnu internationalement par l'ISO (*International Standard Organization*); son objectif est de permettre l'échange de textes sur support électronique.

À l'origine de SGML se trouvent les nombreuses difficultés rencontrées par le monde de l'édition face à l'hétérogénéité des outils informatiques; les professionnels vont s'unir avec un même souci : faciliter l'édition électronique, et ils vont collaborer activement aux divers travaux qui mèneront au standard SGML. Quelles ont été les idées défendues par ces pionniers de SGML?

1. Le balisage spécifique

Un système de traitement de texte manipule deux types d'informations :

— l'information proprement dite, c'est-à-dire le contenu du document : les mots, les figures, les dessins, etc.
— l'information structurelle qui permet de définir le format de présentation du document.

L'information relative à la structure du document est donnée, dans le traitement de texte, sous forme de « balises » insérées aux endroits voulus dans le corps du texte.

Illustrons notre propos par un exemple tiré du traitement de texte *WordPerfect*; sous ce logiciel, un titre accompagné de ses instructions de formatage peut présenter l'aspect suivant :

[Centre][GRAS]The SGML handbook[gras][RT]

Le « décodage » de cet exemple nous apprend :

— que la balise [Centre] positionne le titre au centre de la page;
— que les codes [GRAS] et [gras] indiquent le début et la fin de la zone à imprimer en caractères gras;
— que l'instruction [RT] commande le saut à la ligne suivante.

Malheureusement du point de vue de l'utilisateur du programme, chaque traitement de texte possède un répertoire d'instructions qui lui est propre; l'inconvénient majeur de ce balisage spécifique apparaît lorsqu'un auteur décide de transposer son texte sous un autre éditeur de texte ou sous une autre application telle qu'un programme de recherche documentaire :

— le passage d'un traitement de texte à l'autre nécessitera la reconversion des données ou l'apprentissage d'autres instructions;

— le programme de recherche perdra les instructions de formatage alors que celles-ci peuvent parfois être de première importance : elles établissent, par exemple, la distinction entre le titre et les notes infra-paginales; sans cette information, l'utilisateur effectuera ses recherches tant sur les mots du titre, dont le « poids » informationnel est très important, que sur les mots contenus dans une note.

2. Le balisage générique

On appelle « balisage générique » les instructions de formatage qui ne sont pas limitées à une application unique, à un seul style de formatage ou à un système informatique. Ce concept est introduit au début des années 1970[1] et repose sur deux postulats :

— le balisage doit décrire la structure du document plutôt que ses caractéristiques physiques : ce balisage doit être établi une seule fois et doit suffire pour tous les traitements futurs du document;

— le balisage doit être rigoureux afin qu'il puisse être interprété de façon non ambiguë par un programme ou par une personne.

Quelles propriétés pouvons-nous attendre de ce type de balisage ?

a) La première propriété est de distinguer les éléments significatifs du texte et de les marquer par un nom mnémonique.

Par exemple, le balisage générique des paragraphes se présentera sous la forme :

(deb-para)La première propriété est de ... (fin-para)
(deb-para)Par exemple, le balisage générique ... (fin-para)

Les balises « (deb-para) » et « (fin-para) » marquent respectivement les débuts et fins de chaque paragraphe.

b) Une deuxième propriété est la séparation rigoureuse entre la structure du document et son apparence : le balisage générique spécifie la nature du texte suivant la balise et non le traitement que doit subir ce texte.

Dans l'exemple cité ci-dessus, les balises indiquent que le texte est composé de deux paragraphes : rien n'est dit sur la façon d'imprimer ces paragraphes (marge éventuelle, type de police, hauteur des caractères, etc.).

---

[1] Le concept de « balisage générique », développé par Charles Goldfarb *et al.*, au sein d'*IBM*, est à l'origine du standard « GML » (« *Document Composition Facility Generalized Markup Language* »). GML fut utilisé intensivement durant une décennie. Ce concept fut ensuite repris dans la norme SGML.

c) La dernière propriété découle de ces deux premières caractéristiques : les balises génériques sont indépendantes des systèmes d'édition; elles restent donc inchangées quel que soit le système informatique ou le traitement de texte mis en œuvre.

Publiée en 1986, la norme SGML ISO 8879 [15] est l'aboutissement des recherches menées sur le balisage générique.

Mettre en œuvre SGML

La première étape, dans l'implémentation de SGML, consiste à déterminer la classe générale des documents présentant une même structure : monographies, notices bibliographiques, articles de presse, etc. L'analyse de cet ensemble permettra d'écrire la « définition de type de document ». Le balisage des documents pourra ensuite être réalisé.

1. Le balisage[2]

SGML permet de décrire la structure logique des documents : décrire logiquement un document consiste à déterminer ses composants logiques en dehors de toute considération physique[3]. Par exemple, un livre, considéré sous son aspect logique, est une suite de tomes; chaque tome est une suite de chapitres; chaque chapitre est une suite de sections, chaque section un ensemble de paragraphes, etc. La description physique du même livre distingue les volumes, les pages, les lignes, toutes notions liées au support matériel.

La description logique d'un document, en SGML, s'effectue en insérant dans le texte des balises[4]. Ces balises signalent le début et la fin des éléments logiques contenus dans le document.

Une notice bibliographique, simplifiée, est composée, par exemple, des éléments logiques suivants : l'élément *notice*, l'élément *titre*, l'élément *auteur*,

---

[2] Pour la clarté de l'exposé, nous préférons définir ce qu'est le balisage des documents avant d'expliciter la « définition de type de document ».

[3] SGML permet d'inclure des instructions de formatage — type de fonte, hauteur des caractères, numérotation, etc. — à l'intérieur du texte des documents. Cependant la norme 8879 déconseille vivement l'emploi de ce genre d'instructions car elles vont à l'encontre de l'objectif de SGML : rendre les documents interchangeables et donc indépendants de tout environnement informatique.

[4] Les termes anglais pour « balisage » et « balise » sont, respectivement, « *markup* » et « *tag* ».

l'élément *lieu d'édition* et l'élément *date*. Voici un balisage SGML possible de deux références bibliographiques :

```
<notice>
<titre>The SGML handbook</titre>
<auteur>Charles F. Goldfarb</auteur>
<lieu>Oxford</lieu>
<date>1990</date>
</notice>
<notice>
<titre>Practical SGML</titre>
<auteur>Éric van Herwijnen</auteur>
<lieu>Dordrecht</lieu>
<date>1990</date>
</notice>
...
```

Une balise est composée :
- du nom mnémonique donné à l'élément : *notice, titre, auteur, lieu, date*;
- des délimiteurs de début ou de fin d'élément :
  - une balise de début d'élément est constituée par le nom encadré des délimiteurs *<nnn>*,
  - une balise de fin d'élément est constituée par le nom encadré des délimiteurs *</nnn>*.

2. La définition du type de document (DTD)

Le balisage d'un élément consiste à l'encadrer entre une balise de début et une balise de fin. Les balises d'un document décrivent sa structuration en éléments : elles indiquent les différents éléments qui peuvent être présents dans le document ainsi que l'ordre dans lequel ils apparaissent. Dans notre exemple, une notice bibliographique est composée de cinq éléments : la *notice*, le *titre*, l'*auteur*, le *lieu* d'édition, la *date* et commence toujours par l'élément *titre*.

Les règles définissant la structure d'un document sont indiquées dans une définition de type de document (DTD). Une définition de type de document établit :

— les noms mnémoniques des éléments autorisés dans le document :

> *livre*, *titre1*, *auteur*, *date*, *titre2*, *paragraphe*, etc. sont les noms des éléments structurant le document *livre*.

— les caractéristiques éventuelles de ces éléments :

> L'élément *édition* peut être qualifié par son statut : *1ère* édition, édition *revue et augmentée*, etc.

— les différents constituants d'un élément :

> L'élément *notice* est composé de quatre constituants : le *titre*, l'*auteur*, le *lieu* d'édition et la *date*.

— l'ordre d'apparition des éléments :

> L'élément *notice* commence toujours par le *titre*; celui-ci est suivi par l'*auteur*, etc.

La définition de type de document est donc un schéma de structure du document; elle peut être considérée comme la grammaire générative d'un langage : elle se compose d'un ensemble d'unités syntaxiques — les balises — et d'un ensemble de règles qui déterminent la façon dont ces unités peuvent être mises en œuvre à l'intérieur du document.

La cohérence entre cette syntaxe et les balises SGML introduites dans le texte du document est vérifiée par un «*parser*[5]», c'est-à-dire par un programme d'analyse syntaxique et sémantique. Celui-ci relève les erreurs ou les ambiguïtés introduites au niveau du document, mais aussi dans la définition de type de document.

Quoi de neuf sous SGML?

SGML met en œuvre des balises, tout comme les traitements et les éditeurs de texte. Dans ce cas, quelle est l'originalité de SGML? La nouveauté réside essentiellement dans la signification des balises et dans leur portée :

— Les codes utilisés dans les traitements de texte courants fournissent essentiellement une instruction relative à l'aspect final du document imprimé; elles ne sont interprétables que par le traitement de texte qui les a générées;

— Les balises conçues sous SGML mettent en évidence la structure hiérarchique du document : celui-ci est perçu comme un arbre aux nombreuses ramifications; l'objectif de SGML est de partir du sommet de l'arbre, qui est le document lui-même, et de décrire toutes ses branches; cette description reste valable, quelle que soit la destination finale du document.

---

[5] Le «*parser*» SGML est un programme, ou une suite de programmes, qui permet de contrôler la validité du document balisé en SGML en fonction des règles définies dans la DTD. Certaines techniques, décrites dans la norme, permettent en outre de réduire le nombre de balises à insérer dans le corps du document. Le rôle du *parser* est aussi de reconstituer les balises omises. Le résultat du «*parsing* est un document logiquement parfait et dont le balisage est complet.

Reprenons notre exemple, la liste de références bibliographiques :

— sous *WordPerfect*, les notices pourront être codées de la façon suivante :

```
[Tab][SOUL]The SGML handbook[soul][RT]
[Tab][GRAS]Charles F. Goldfarb[gras][RT]
[Tab]Oxford[RT]
[Tab]1990[RT]
[Tab][SOUL]Practical SGML[soul][RT]
...
```

— voici la même liste balisée en SGML :

```
<notice>
<titre>The SGML handbook</titre>
<auteur>Charles F. Goldfarb</auteur>
<lieu>Oxford</lieu>
<date>1990</date>
</notice>
<notice>
<titre>Practical SGML</titre>
...
```

WordPerfect ne donne aucune indication concernant la structure logique du document; celle-ci est comprise dans les instructions de formatage; ces dernières sont propres au traitement de texte et ne peuvent être interprétées par d'autres programmes.

SGML met en évidence le rôle joué par les caractères situés entre les balises de début et de fin de zone : il s'agit d'un titre, d'un auteur ou d'une date. SGML ne fournit, par exemple, aucune information concernant le format d'impression du titre. Ces instructions, spécifiques à un environnement informatique précis, seront développées dans un ou plusieurs programmes d'application...[6]

[6] D'aucuns objecteront que cet apport de SGML n'en est pas un dans la mesure où il a été précédé par d'autres langages mettant en oeuvre le concept de balisage générique. Par rapport à ces derniers, SGML possède l'atout majeur d'être le premier langage de balisage générique reconnu par l'ISO.

L'application SGML

Quelle est la différence entre SGML tel que nous l'avons brièvement présenté ci-dessus et ce qui est appelé « l'application SGML » ?

SGML est un ensemble de règles énoncées dans la définition de type de document. Pour rappel, ces règles sont prévues :

— pour définir les différents types de documents : document de type *monographie*, de type *article de revue*, de type *journal*, etc.;

— pour identifier les éléments logiques de ces documents au moyen des balises descriptives des éléments.

Une application SGML met en relation les descriptions logiques du document balisé et un usage précis tel l'édition du document. Le passage entre balisage logique et application SGML est assuré par un programme capable d'extraire et de traiter l'information codée en SGML.

Une application SGML pourrait, par exemple, remplacer toutes les balises SGML contenues dans un fichier par des commandes de composition[7] :

UTILISER Baskerville Italic 10 pt
The SGML handbook
ALLER ligne
UTILISER Baskerville bold 10 pt
ALLER ligne
Charles F. Goldfarb
UTILISER Baskerville Roman 10 pt
Oxford
...

L'impression de ce document se présentera sous la forme :

The SGML handbook
Charles F. Goldfarb
Oxford
...

[7] Ces commandes sont pure invention de notre part.

SGML et l'édition

Les éditeurs ont fait un grand pas en avant le jour où ils ont obtenu de leurs auteurs la fourniture des textes sur support électronique. Le processus de publication a ainsi été grandement accéléré, mais les obstacles demeurent...

1) Du point de vue de l'auteur

   L'auteur, décidé à s'équiper en informatique, a un premier souci : choisir un ordinateur et une imprimante dans la vaste gamme des produits qui lui sont proposés. Le choix de l'un et l'autre modèle impliquera l'acquisition de logiciels compatibles avec la configuration prévue et limitera presque nécessairement les opérations permises. Cependant, le problème surgit réellement lorsque les exigences de l'éditeur ne correspondent pas à l'environnement informatique de l'auteur : imposer la saisie du texte sous *WordPerfect* version 5.1 alors que l'auteur possède la version antérieure ou ne dispose pas de ce traitement de texte... La solution ? demander à l'éditeur de réduire ses exigences ou trouver la configuration informatique adéquate.

2) Du point de vue de l'éditeur

   Le principal problème de l'éditeur réside dans la conversion des documents reçus. Au pire, l'éditeur reçoit une version papier de son auteur; le manuscrit doit être saisi sur support électronique, ce qui signifie une perte de temps... et d'argent. Au mieux, le texte est envoyé selon le format requis par l'éditeur : le texte doit être mis en forme en respectant les souhaits de son auteur ainsi que le style de la collection ou de la revue dans lequel le texte doit être inséré.

Face à ce qui ressemble souvent à un « parcours du combattant », les éditeurs ont réclamé un outil standard afin de faciliter la préparation et la publication des documents électroniques. De cette demande est née la norme SGML.

Pour le monde de l'édition, les avantages liés à ce standard sont essentiellement économiques :

— facilité de communication entre systèmes hétérogènes : l'auteur saisit le texte de son article à domicile, en insérant les balises SGML en fonction de la structure logique de sa publication; l'éditeur réceptionne le texte au format éditeur et met en œuvre les programmes applicatifs de composition;

— apprentissage, par les différents protagonistes, d'un seul type de procédures : l'auteur insère les balises SGML au moyen d'un éditeur SGML[8];

[8] Un certain nombre de programmes « éditeurs SGML » existent sur le marché; ils sont portables sur les environnements micro-informatiques les plus courants. Leurs tâches principales sont : faciliter le travail d'insertion des balises dans le texte du document et contrôler que le balisage en cours respecte la définition de type de document.

— facilité de mise à jour des données : les informations sont indépendantes de l'ordinateur et de l'application; une modification du fond ou de la forme ne signifie pas nécessairement retour à l'auteur, re-saisie du texte ou recomposition.

Les éditeurs ont été parmi les premiers à implémenter le standard SGML. En Europe, l'Office des Publications Officielles des Communautés Européennes est la première institution à avoir adopté SGML [21] : le codage, en SGML, du *Journal Officiel* a permis la création d'une base de données portant sur la législation européenne[9].

Les autres possibilités de SGML

Les applications permises par le standard ISO s'étendent bien au-delà du secteur de l'édition. Néanmoins, l'outil SGML est récent : peu d'applications sont réellement fonctionnelles et les possibilités multiples du standard n'ont pas encore pu être exploitées. Voici quelques voies déjà explorées.

1. La structuration de textes en vue de la recherche d'informations

Le texte d'un ouvrage ou d'un article contient des informations dont la portée est différente selon que celle-ci se situe dans une note infra-paginale ou en tête de chapitre. Seule l'apparence physique du document permet d'établir la distinction entre la zone « note » ou la zone « chapitre » : par exemple, la note sera imprimée en bas de page, dans une fonte plus petite, alors que le chapitre sera précédé d'un titre et commencera par une indentation.

Les systèmes traditionnels de recherche d'information ne conservent pas la structure des textes qu'ils stockent. Ils considèrent le document comme un ensemble de mots classés dans différentes rubriques — rubriques *auteur*, *titre*, *date*, *texte*, etc. — et la base de données documentaire est vue comme une collection de documents non structurés[10] : par exemple, l'indication que « *le terme "normalisation" appartient au titre du premier chapitre* » est perdue lors du stockage des documents dans la base de données.

---

[9] À cette occasion, la *Sobemap*, société belge, a développé le premier « *parser* » commercial.

[10] Certains logiciels, tels *Stairs* ou *Status*, permettent une première structuration du document en distinguant, par exemple, les « paragraphes » ou les « sections » à l'intérieur du texte.

Comment concilier les avantages des outils automatisés de recherche d'information tout en préservant la structure apparente des textes enregistrés dans la base de données ?

Une équipe universitaire de l'Ontario a apporté un premier élément de réponse en développant un prototype de stockage et de recherche de textes structurés : *Maestro*[11] [3][17]. L'objectif de *Maestro* est d'exploiter la structure des textes afin d'établir des stratégies de recherche basées sur le contexte. Pour réaliser cet objectif, *Maestro* « capture » la structure logique implicite des documents balisés en SGML. Voici le processus de traitement d'un document sous Maestro :

— le texte est balisé en SGML, puis « parsé » par XGML[12] ;
— le « *parsing* » produit une représentation de la structure du document sous forme d'arbre;
— une procédure interne au système passe en revue cet arbre syntaxique : elle reconstitue le document original et construit une table; dans cette table, chaque mot du texte est classé en fonction de sa localisation dans le document;
— le document « plat », c'est-à-dire sans ses informations structurelles, est ajouté à la base de données et l'arbre généré par le parser est stocké comme un « document » séparé.

La base de données ainsi alimentée permet des recherches affinées sur le contexte d'un mot telles que :

```
trouver les documents dont un ou plusieurs chapitres ont
un titre contenant le mot "Simenon" et dans lesquels
la première phrase des paragraphes composant ce ou ces
chapitres présente le mot "oeuvre".
```

Le système résoud cette requête en plusieurs étapes :

— pour chaque document possédant le terme *Simenon*, il détermine les positions de celui-ci à l'intérieur du document;
— il vérifie dans la table si ces positions correspondent à l'emplacement « titre de chapitre » ;
— il recherche tous les documents contenant le mot « œuvre » ;
— il établit l'intersection entre le résultat obtenu sur « œuvre » et le résultat précédent;
— pour chaque document retenu, il définit la position de « œuvre » à l'intérieur du texte;

---

[11] *Management Environment for Structured Text Retrieval and Organization.*

[12] XGML est un « *parser* » américain.

— il contrôle que cette position correspond à la première phrase des paragraphes;
— il localise, pour chaque document finalement sélectionné, les chapitres pertinents.

Ainsi conçu, le système *Maestro* fournit des possibilités de recherche plus puissantes et plus souples que celles habituellement disponibles sur les systèmes de recherche de textes non structurés[13].

2. La production d'index à partir de termes d'index balisés dans le texte lui-même

Le projet développé dans *Maestro* s'applique parfaitement pour un texte au format relativement simple. Mais que faire lorsque le document contient du texte, des figures, des formules chimiques, des tableaux, des images, etc. c'est-à-dire lorsqu'il mêle à la fois caractères alphanumériques, graphiques et données numérisées.

Aux États-Unis, un projet sponsorisé par l'OCLC[14] a trouvé une solution à cette problématique [14]. L'objectif à la base de cette recherche était de mettre à la disposition des scientifiques et des utilisateurs fréquentant la bibliothèque une encyclopédie de technologie chimique[15]. La difficulté résidait dans l'hétérogénéité du contenu de cette littérature chimique : symboles et caractères divers, équations complexes, tables étendues, dessins et photographies de couleur.

Le choix des chercheurs s'est porté sur SGML. La création de la base de données s'effectue de la façon suivante :
— en fonction de son contenu, l'ouvrage encyclopédique subit deux traitements différents :
  – le texte est saisi en caractères ASCII[16] et balisé en SGML;
  – les images, photos et autres données graphiques sont scannées;
— la mise en œuvre de programmes applicatifs permet de générer un index reprenant les informations textuelles;

---

[13] La société française *AIS* (*Advanced Information Systems*) a annoncé, en novembre 1990, le lancement d'un nouvel outil de recherche : *SGML/Search*. Ce produit devrait «convertir les documents balisés en SGML en bases de données» et permettre des recherches de type contextuel. Nous attendons une documentation à propos de ce logiciel.

[14] *Online Computer Library Center.*

[15] La *Kirk-Othmer Encyclopedia of Chemical Technology.*

[16] L'application a été développée sur un micro-ordinateur tournant sous le système opératoire MS-DOS.

— les données numérisées sont organisées et mises en relation avec le texte.

Le résultat de ce travail est un fichier informatique autorisant diverses manipulations telles que :

— le repérage des mots du texte par l'insertion de balises SGML et l'indexation complète de ces termes;
— la conversion des tables et équations, exclues de l'index, dans le format proposé par TEX[17], un marqueur de textes scientifiques;
— la recherche «*full-text*» et l'affichage, en réponse, des articles accompagnés des tableaux, figures et autres illustrations correspondantes;
— la mise à jour aisée des articles;
— la composition et l'impression du texte selon le style souhaité.

3. L'exploitation des archives électroniques

À l'heure de l'informatique, les archives ne sont plus seulement constituées d'ouvrages poussiéreux ou de liasses de papiers personnels : elles se sont enrichies des données enregistrées sur support électronique. La conservation de ces archives peut être rentabilisée, d'un point de vue commercial, si elle est réalisée dans de bonnes conditions [24].

Le but de l'archivage est de préserver un document de valeur dans l'éventualité d'un usage futur. L'intégration d'un bon codage générique — SGML en l'occurrence — peut apporter une valeur ajoutée aux archives électroniques [26]. De par sa nature, le codage générique autorise le recyclage de produits considérés comme « hors-circuit » vers d'autres objectifs. « Redonner vie » à un produit dépassé peut être imaginé de différentes façons :
– création d'une base de données *on-line*; – élaboration de CD-ROM; – nouvelles versions imprimées de tout ou partie des archives.

Par exemple, pour valoriser l'*Oxford English Dictionary*, ses éditeurs ont décidé d'exploiter les outils électroniques [19]; les seize volumes que compte actuellement le dictionnaire ont été saisis sur support informatique puis balisés en SGML. Les archives électroniques du dictionnaire restent intactes, mais l'intégration de SGML permet de proposer l'ouvrage en ligne, en interrogeant une banque de données ou un CD-ROM. D'autres avantages peuvent être relevés telle la mise à jour rapide des articles du dictionnaire, des éditions supplémentaires introduisant une présentation plus moderne ou exploitant de nouveaux

[17] C'est le logiciel TEX qui est utilisé pour la composition de cette revue (N.D.É.).

supports. Forte de cette première expérience en SGML, la maison d'édition Oxford University Press collabore actuellement avec la maison d'édition française Hachette afin de produire un nouveau dictionnaire bilingue anglais/français.

D'autres recherches, en cours de développement, méritent d'être mentionnées tels le projet PERSEUS[18] étudié par le département de philologie classique à la Harvard University de Cambridge, aux États-Unis [4]. L'objectif de PERSEUS est de construire une vaste banque de données multimédia : celle-ci devrait assister les étudiants, les chercheurs et les professeurs dans l'étude du monde grec classique. La base de données devrait contenir des textes de littérature antique (textes grecs et traductions anglaises), un dictionnaire grec, des images, des plans de sites antiques, etc. Le balisage des données en SGML permet de garantir la « portabilité » des fichiers vers d'autres applications.

Autre recherche, celle menée par la Faculté des Arts de l'Université d'Oslo [4]. Prévu pour une durée de six ans, le projet « Documentation Project » vise à créer, sur support électronique, un corpus de textes norvégiens aussi complet que possible. La base de données lexicale devrait être créée à partir d'une large collection de fiches reprenant des termes dialectaux. SGML serait utilisé comme langage de balisage.

SGML, ODA et les autres ?

Qui aborde SGML se doit de mentionner ODA[19], une norme relative, elle aussi, à la description des documents [7].

La norme ODA est souvent considérée, à tort, comme compétitive de SGML. ODA est une norme internationale permettant l'échange des documents dans un environnement bureautique. À la différence de SGML, ODA donne une description précise et détaillée de l'interprétation à donner au document par le destinataire : celui-ci est capable de recréer exactement la présentation du document (le *layout*) telle qu'elle a été prévue et conçue sur le système dont il est issu.

Bien que différentes, la co-existence de ces deux normes internationales pose à nouveau le problème de compatibilité : opter pour l'une implique, presque

---

[18] Pour une description de ce projet, on se reportera à l'article Fr. CHARPIN, *Les programmes PERSEUS et ITER*, dans *Revue. Informatique et Statistique dans les Sciences humaines*, 27 (1991), pp. 97–101 (N.D.É.).

[19] La norme ODA est une norme ISO publiée en 1988 («ISO 8613 – Office Document Architecture»).

obligatoirement, le rejet de l'autre... C'est pourquoi, des chercheurs se sont réunis afin d'étudier la possibilité de « réconcilier » ODA et SGML : aujourd'hui, le dialogue entre les deux standards est encore loin d'être une évidence...

Le mot de la fin...

La pierre d'achoppement du monde informatique, c'est l'individualisme et le manque de communication. Le standard SGML tente d'apporter une première réponse à ce problème majeur. S'il faut en voir les premières applications et s'il faut en croire les premiers bénéficiaires, l'essai est concluant.

Néanmoins, tout n'est pas rose non plus dans le monde SGML. En effet, beaucoup d'utilisateurs potentiels de SGML sont découragés par le côté hermétique de la norme — le standard est très complexe — et les difficultés pratiques de sa mise en œuvre : l'analyse des informations à structurer, la définition des balises et des règles de structuration, le développement de programmes de conversion sont affaires de spécialistes. Ainsi, implémenter SGML entraîne souvent une surcharge de traitements considérable. De cette constatation en découle une seconde : SGML est une norme à grande échelle; elle se situe au niveau de l'entreprise, de l'organisation et non de l'individu...

Alors, SGML, « une norme pour les autres mais pas pour moi ? ». Peut-être ne sera-ce pas vous, mais les autres, qui prendront l'initiative d'introduire SGML. Mais n'en serez-vous pas, un jour, les heureux bénéficiaires...?

Orientation bibliographique

[1] J. ANDRÉ, R. FURUTA, V. QUINT, *Structured documents*, Cambridge University Press, Cambridge, 1989.

[2] S. ARNOLD, *Electronic Publishing*, in *New Zealand Libraries*, vol. 45, n° 12, 1988, pp. 275–278.

[3] D.T. BARNARD, D. HAMILTON, M. LEVINSON, I.A. MACLEOD, *SGML Documents and Non-linear Text Retrieval*, in *RIAO 91. Conference Proceedings. Intelligent Text and Image Handling*, Universitat Autonoma de Barcelona, Barcelona, Spain — April 2–5, 1991, pp. 226–243.

[4] D.T. BARNARD, R. COVER, N. DUNCAN, *Bibliography on SGML (Standard Generalized Markup Language) and related issues. Technical Report 91–299*, s.l., février 1991.

[5] M. BLAKE, *Practices of printers in relation to electronic archiving*, in *The Electronic Library*, vol. 8, n° 4, 1990, pp. 239–241.

[6] J. BLEEKER, J.G. KIRCZ, *The use of relational databases for electronic and conventional scientific publishing*, in *Journal of Information Science*, vol. 13, 1987, pp. 75–89.

[7] C. BORMANN, U. BORMANN, *Standards for open document processing : current state and future developments*, in *Computer Networks and ISDN Systems*, vol. 21, 1991, pp. 149–163.

[8] N. BRADLEY, *SGML concepts*, in *ASLIB Proceedings*, vol. 44, n^{os} 7/8, 1992, pp. 271–274.

[9] M. BRYAN, *An author's guide to the standard generalized markup language*, Addison-Wesley ed., Reading, Mass., 1988.

[10] M.C.S. BUCKINGHAM, *At the coalface : Author-Typesetter-Publisher links*, in *Electronic Publishing : the new way to communicate. Proceedings of the symposium organized at the Commission of the European Communities*, Luxembourg, 5–7 Nov. 86, pp. 143–154.

[11] D. VIGNAUD, *SGML. Application à l'édition française*, Éditions du Cercle de la Librairie, Paris, 1989.

[12] Ch. F. GOLDFARB, *The SGML handbook*, Clarendon Press, Oxford, 1990.

[13] Ch. M. GOLDSTEIN, *Online reference works and full text retrieval*, in *National Online Meeting 1989. Proceedings of the 10th National Online Meeting*, New York, 9–11 May 1989, pp. 171–177.

[14] Th.B. HICKEY, *Using SGML and TEX for an interactive chemical encyclopedia*, in*National Online Meeting 1989. Proceedings of the 10th National Online Meeting*, New York, 9–11 May 1989, pp. 187–195.

[15] *Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML). ISO 8879–1986(E)*, International Organisation for Standardization, 1986.

[16] V.M. KRÜGER, *Technische Publikation: Strategien und Systeme*, in *Nachrichten für Dokumentation*, vol. 41, n° 5, 1990, pp. 285–289.

[17] I.A. MACLEOD, *Storage and retrieval of structured documents*, in *Information Processing & Management*, vol. 26, n° 2, 1990, pp. 197–208.

[18] O. LAURENT, *Approche grammaticale des structures de documents*, Mémoire d'informatique, Université Libre de Bruxelles, 1990–1991.

[19] J.D. PAINTER, *Marking up the Dictionary*, in *Information Media & Technology*, vol. 21, n° 2, 1988, pp. 72–74.

[20] B. ROBINSON, G. WU, *Applications of SGML*, in *Online Information 90. 14th International Online Information Meeting. Proceedings*, London 11–13 December 1990, Oxford — New Jersey, 1990, pp. 75–82.

[21] *SGML. Briefing note (n^{os} 1–16)*, published by the Commission of the European Communities — Directorate — General XIII-B. Telecommunications, Informations, Industries and Innovation.

[22] *Un regard neuf sur SGML : les idées toutes faites évoluent*, in *Normatique. La normalisation des technologies de l'information*. Numéro spécial, Paris, mai 1992.

[23] E. VAN HERWIJNEN, *Practical SGML*, Kluwer Academic Publishers, Dordrecht, 1990.

[24] R. WILLIAMSON, *Electronic Text Archiving*, Oxford, Elsevier, 1988.