Le langage XML

JON BOSAK • TIM BRAY

Le réseau Internet s'est considérablement développé grâce à la mise au point de systèmes «hypertexte» qui gèrent les images, les textes, les sons. Le langage XML accentuera ce développement.

Quelques indices nous suffisent pour reconstruire une information : en jetant un coup d'œil à cette page, et en repérant des mots imprimés en gros caractères, suivis de colonnes en petits caractères, nous déduisons que nous avons sous les yeux le début d'un article. De même, un simple coup d'œil sur une liste de produits d'épicerie montre que celle-ci est une liste de courses. En consultant quelques rangées de chiffres, on identifie un relevé de compte bancaire.

Les ordinateurs n'ont pas cette capacité : on doit leur indiquer précisément la nature des objets qu'on leur fait manipuler, les liens entre ces objets et ce qu'ils sont censés en faire. Le langage XML (*eXtensible Markup Language*, soit «langage de balisage extensible») est conçu justement pour rendre l'information autodescriptive. Ce changement de la communication entre les ordinateurs, simple en apparence, pourrait étendre l'utilisation de l'Internet à nombre d'activités humaines, au-delà de la simple transmission d'information. Depuis la création de XML, au début de 1998, par le Consortium W3C *(World Wide Web Consortium)*, il s'est répandu comme une traînée de poudre dans les milieux scientifiques et industriels.

Cet engouement s'explique par l'espoir que XML comblera certaines des principales lacunes du réseau Internet. L'une d'elles est la lenteur : sur le réseau Internet, l'information est censée circuler à la vitesse de la lumière, mais, en pratique, elle arrive aux utilisateurs à la vitesse de l'escargot. En outre, bien qu'une quantité phénoménale d'information soit disponible, la recherche de l'information adéquate est souvent fastidieuse.

Ces deux problèmes sont en grande partie dus à la nature du langage principal du réseau, le langage HTML (l'acronyme de *Hypertext Markup Language*, soit «langage de balisage hypertexte»). Bien que celui-ci soit le langage de publication électronique le plus populaire jamais inventé, il reste superficiel : il décrit seulement comment les programmes de navigation sur le réseau doivent organiser le texte, les images et les boutons sur les pages qu'ils affichent. Cette superficialité de HTML le rend facile à apprendre, mais cette simplicité a un prix.

Notamment, avec HTML, on crée difficilement des sites qui ne se limitent pas à envoyer des documents à ceux qui en font la demande, tels des télécopieurs améliorés. Les particuliers et les entreprises veulent des sites qui enregistrent les commandes des clients, qui transmettent des dossiers médicaux ou, même, qui gèrent des usines ou des instruments scientifiques à l'autre bout du monde. Le langage HTML n'est pas conçu pour de telles tâches.

Aujourd'hui, un médecin peut récupérer un dossier médical à l'aide d'un navigateur, mais il ne peut ensuite l'envoyer électroniquement à un collègue spécialiste afin qu'il l'intègre directement à la base de données de son hôpital. Son ordinateur ne sait pas quoi faire de l'information, qui n'est que du <H1>bla bla </H1> <BOLD>bla bla bla<p>. Le problème avec le «Ce Que Vous Voyez Est Ce Que Vous Obtenez» (la traduction française de *What You See Is What You Get* ou WYSIWYG), c'est que ce que vous voyez n'est que ce que vous obtenez.

Les étiquettes entre crochets, dans l'exemple précédent, sont des balises. Le langage HTML ne possède pas de balise qui signale une réaction allergique à des médicaments, par exemple ; sa raideur est une autre de ses limites. L'adjonction d'un nouveau type de balises impose des démarches administratives qui peuvent prendre des années, si bien que peu s'y aventurent. Pourtant, chaque application, et pas seulement l'échange de dossiers médicaux, nécessite ses propres balises.

Cette raideur explique également la lenteur actuelle des bibliothèques en ligne, des catalogues de vente par correspondance ou des divers sites interactifs. Quand on veut modifier les quantités commandées ou le mode d'acheminement des marchandises, le serveur doit renvoyer une nouvelle page, dont du texte et des graphiques déjà reçus. Pendant ce temps, votre puissant ordinateur attend bêtement, parce qu'il ne connaît que les <H1> et les <BOLD>, et non les prix et les options de transport.

De là résultent également les recherches frustrantes sur le réseau. Parce qu'il n'existe pas de moyen de signaler qu'une donnée est un prix, vous ne pouvez pas utiliser les informations tarifaires dans vos recherches.

Comment faire du neuf avec du vieux ?

En théorie, le problème serait résolu si des balises décrivaient le contenu de l'information, et non ce à quoi elle ressemble. Par exemple, au lieu de classer les différentes composantes d'une commande d'après les rubriques «caractères gras», «paragraphe», «ligne» et «colonne» (ce que fait le langage HTML), on les dénommerait «prix», «taille»,

«quantité» et «couleur». De la sorte, un programme identifierait un document contenant ces rubriques comme étant une commande, et il le gérerait de manière souple : il pourrait l'afficher, le rentrer dans un système de gestion ou vous faire livrer une nouvelle chemise demain.

Dès 1996, un groupe de travail comprenant une dizaine de membres du Consortium W3C a exploré une telle solution. L'idée, bien que puissante, n'était pas entièrement originale : depuis des générations, les éditeurs annotent les manuscrits pour guider les typographes. Ce système de balisage (*markup* en anglais) s'est progressivement formalisé jusqu'à ce qu'en 1986, après des années de travail, l'Organisation des normes internationales (ISO) approuve un système pour la création de nouveaux langages de balisage.

Nommé SGML (pour *Standard Generalized Markup Language*, soit «langage de balisage généralisé standard»), ce métalangage (langage de description de langages) a fait la preuve de son utilité dans l'édition. Le langage HTML lui-même a été défini à partir du SGML. Le seul inconvénient du SGML est qu'il est trop général, plein d'astuces destinées à minimiser les frappes, à une époque où chaque octet avait un coût non négligeable. Sa complexité est excessive pour les programme de navigation.

Notre groupe de travail a créé XML en épurant le SGML pour obtenir un métalangage plus simple. XML est un ensemble de règles que chacun peut suivre pour créer un langage de balisage. Les règles de XML garantissent qu'un seul programme simple et compact, souvent nommé analyseur, est capable de traiter tous ces nouveaux langages.

Reprenons l'exemple du médecin qui veut envoyer un dossier médical à un spécialiste. Si le corps médical utilise XML pour fabriquer un langage de balisage des dossiers médicaux (plusieurs groupes étudient ce cas précis), le courrier électronique d'un médecin signalera une réaction à des médicaments par un code tel que <patient> <nom> Moreau </nom> <allergie médicamenteuse> pénicilline </allergie médicamenteuse> </patient>. Un programme très simple permet à un ordinateur de comprendre cette notation médicale standard et d'ajouter cette statistique vitale à une base de données.

Comme le langage HTML, qui permet à tous les utilisateurs d'ordinateurs de lire les documents sur l'Internet, XML pourrait être le nouvel espéranto informatique. De surcroît, à la différence de la plupart des formats de données informatiques, les

Illustrations de Brucie Rosch

1. LE LANGAGE XML est un pont entre des systèmes informatiques hétérogènes. Il permet la recherche d'informations, l'échange de données scientifiques, de produits commerciaux et de documents multilingues avec une vitesse et une facilité accrues.

annotations XML ont un sens pour les humains, puisqu'elles sont composées de texte ordinaire.

Le potentiel unificateur de XML résulte de quelques règles bien choisies. L'une d'entre elles stipule que les balises vont presque toujours par paires. Comme les parenthèses, elles encadrent le texte auquel elles se rapportent. Comme des guillemets, les paires de balises peuvent être emboîtées.

La règle d'emboîtement simplifie les documents XML en leur conférant une structure d'arbre. À la manière d'un arbre généalogique, chaque graphique ou morceau de texte du document représente un parent, un enfant ou un frère ; les parentés sont sans ambiguïté. Les arbres ne peuvent pas représenter n'importe quelle information, mais ils représentent la plupart des types d'information nécessaires au fonctionnement des ordinateurs. De surcroît, les arbres sont informatiquement très pratiques. Si un relevé de compte bancaire se trouve sous forme d'un arbre, on écrit très simplement un petit programme qui ordonne les transactions ou qui affiche seulement les chèques tirés avant une certaine date.

Le potentiel unificateur de XML s'accroît de son association avec Unicode, un système de codage de caractères qui autorise le mélange de textes dans les principales écritures de la planète. En

Bases de données XML

Peut-on considérer un répertoire de quelques milliers ou millions de documents XML comme une base de données et y retrouver des informations avec la même efficacité qu'avec des systèmes classiques de gestion de bases de données? Imaginons que, demain, des milliers de musées dans le monde publient leurs catalogues de collections et d'expositions en XML et placent ces documents sur des serveurs Web. Si l'on demande à un moteur de recherche : «Existe-t-il un portrait de Blaise Pascal, et si oui qui en est l'auteur?», obtiendrons-nous la réponse «Philippe de Champaigne» sans qu'elle soit noyée sous un déluge d'informations non pertinentes ayant trait à tous les Pascal de l'Univers, ou, au moins, à toutes les œuvres, à tous les commentaires ou biographies écrits à propos de Blaise Pascal? Avec son système de balisage, XML peut grandement faciliter la recherche, même si, à elle seule, cette norme ne résout pas toutes les difficultés de la recherche de l'information en ligne.

Pour améliorer la recherche d'information sur le Web, plusieurs approches non exclusives sont envisageables. L'une d'elles, mentionnée par J. Bosak et T. Bray, est celle des «métadonnées», la transposition au monde numérique des pratiques des bibliothécaires et des documentalistes : à chaque document, on associe une «fiche signalétique». Ces fiches peuvent être rangées dans un catalogue (une base de données) d'où elles peuvent être rapidement retrouvées. Cette approche, appliquée au Web, ne convainc pas, loin s'en faut : nous ne sommes pas tous des documentalistes et, s'il fallait systématiquement créer et gérer des fiches descriptives précises pour chaque document numérique mis sur le réseau, le coût et la complexité seraient rédhibitoires. Or, si ces fiches sont trop simplifiées, elles ne permettent plus cette précision dans la recherche d'information qui est l'objectif visé.

Une variante consiste à indexer les documents par les termes figurant dans certains éléments bien choisis des documents eux-mêmes : les métadonnées descriptives doivent être trouvées dans les documents. Si toutes les informations pertinentes figurent explicitement dans les documents et y sont identifiées par des balises, il n'est nul besoin de créer une fiche descriptive avec un nom d'auteur, un titre, une date de publication, des mots clefs, etc.

Le problème devient alors de partager des systèmes de balisage communs pour ces éléments devant servir à l'indexation des documents, tout en sachant que ces balises doivent aussi être utilisées directement dans les documents les plus divers, aux structures les plus variées et écrits dans des langues différentes. Il reste aussi à inventer des méthodes de recherche qui exploitent ces index créés à partir de ces fragments de documents et qui restent efficaces lorsque le nombre de documents indexés devient très grand.

Une autre approche consiste à structurer différemment l'hypertexte que forme le Web. Aujourd'hui les liens relient entre eux des fichiers, et, pour un ordinateur, un lien du Web signifie seulement une adresse physique de fichier sur le réseau. Si vous voyez dans une page Web la phrase «Vous pouvez contacter l'auteur…», avec le mot auteur mis en valeur par une couleur et un soulignement, vous savez qu'il s'agit d'un lien qui matérialise la relation qui existe entre le document que vous lisez et son auteur, représenté par son adresse de courrier électronique ou sa page d'accueil. Pourtant vous ne pouvez pas demander à un serveur Web de vous retrouver l'adresse électronique des auteurs des pages hébergées : l'information existe, mais sous une forme inexploitable pour un programme. Or, XML permet de typer les liens, c'est-à-dire de spécifier leur sémantique, soit directement, soit en les décrivant comme des relations entre documents dans le système descriptif du langage.

Si les auteurs utilisent des liens qui correspondent à des relations prédéfinies (être auteur de, être une biographie de, être localisé à, etc.), le Web se transforme en une sorte de «graphe conceptuel», et il permet au moins la recherche des groupes de documents par l'exploitation de ces relations explicites définies entre eux.

Ces approches et plusieurs autres sont activement explorées par les spécialistes des bases de données et de la recherche d'information en ligne. Il est d'ailleurs probable qu'aucune méthode ne satisfera à elle seule tous les besoins et que l'on utilisera une panoplie de méthodes complémentaires. Une chose est sûre en tout cas : l'arrivée de XML a relancé et orienté la recherche dans ce domaine et sera la source de grands progrès en matière de recherche d'information en ligne.

Alain MICHARD, INRIA, Rocquencourt