

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

HTML, comme dans la plupart des programmes de traitement de texte, un document est en général dans une seule langue : français, japonais, arabe... Si le programme ne sait pas lire les caractères de la langue de l'utilisateur, ce dernier ne peut pas utiliser le programme. Pis encore, les logiciels destinés à être utilisés à Taiwan sont souvent incapables de lire les textes issus de la Chine continentale, à cause d'incompatibilités du codage des caractères. Au contraire, un programme qui comprend correctement le XML peut faire face à n'importe quelle combinaison de ces jeux de caractères. XML permet l'échange d'information non seulement entre des systèmes informatiques différents, mais aussi au travers des barrières nationales et culturelles.

La fin de l'attente

L'avènement de XML devrait accélérer les échanges sur le réseau Internet. Aujourd'hui, les ordinateurs connectés au réseau, quelle que soit leur puissance, récupèrent des formulaires, les remplissent et leur font faire la navette jusqu'à un serveur, jusqu'à l'aboutissement de la requête. L'information structurelle et sémantique qu'ajoute XML permet aux ordinateurs de faire une bonne partie du traitement localement, sans intervention des serveurs distants, qui se trouvent soulagés d'autant. En outre, le trafic sur le réseau sera également réduit.

Pourquoi ce gain de rapidité? Imaginez que vous vous connectiez au site Internet d'une agence de voyages afin de connaître tous les vols Paris-New York du 11 août. Vous recevez une liste plus longue que ce que votre écran peut afficher. Comment préciser votre choix? En spécifiant l'heure approximative de départ, la gamme de prix ou la compagnie aérienne que vous désirez emprunter. Toutefois, pour obtenir cette liste écourtée, vous devrez envoyer une nouvelle requête au serveur de l'agence de voyages et attendre la réponse. Si la longue liste de vols avait été envoyée en XML, l'agence de voyages aurait pu vous fournir, en même temps que les caractéristiques des vols, un petit programme (en langage Java) que vous auriez pu utiliser sur votre machine pour faire un tri à votre convenance en quelques microsecondes, sans solliciter à nouveau le serveur (voir la

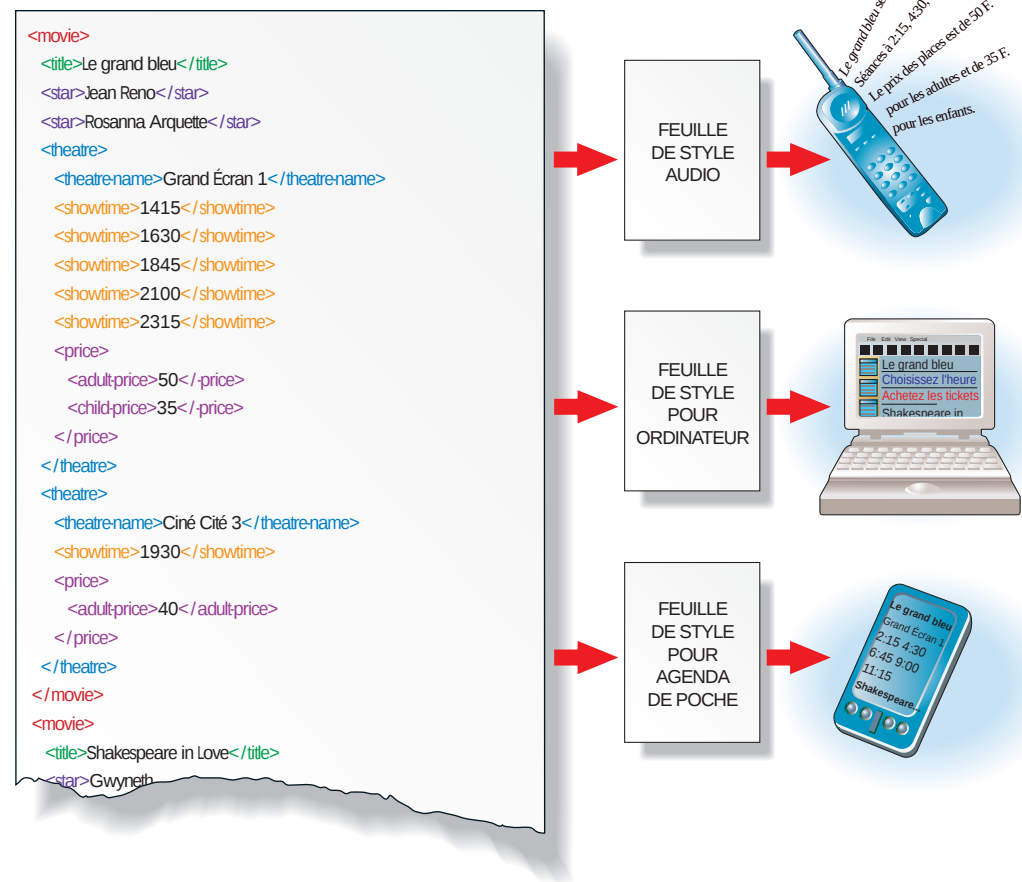
figure 3). Ce gain de temps, multiplié par plusieurs millions d'utilisateurs, sera la cause d'une amélioration notable du réseau Internet.

Plus l'information circulant sur le réseau sera étiquetée avec des balises XML spécifiques à chaque corps de métier, plus on trouvera facilement des informations précises. Aujourd'hui une recherche d'«offres d'emploi d'ingénieur système» récupère beaucoup de publicités, mais peu d'offres, parce que la plupart sont dans les services de petites annonces des sites des divers journaux, hors de portée des moteurs de recherche. L'Association américaine de la presse écrite met au point, pour les petites annonces, un langage de balisage fondé sur XML qui rendra de telles recherches plus fructueuses.

L'efficacité des recherches augmentera encore quand on utilisera une technique employée par les bibliothécaires. Depuis longtemps, ces derniers savent que le meilleur moyen de trouver une information rapidement n'est pas de rechercher l'information elle-

même, mais plutôt des ensembles de données plus spécifiques, qui orientent vers les sources utiles. Ces informations sur l'information, dont les catalogues de bibliothèque sont un bon exemple, sont des métadonnées.

Depuis le début du projet XML, on envisage la création d'un standard pour les métadonnées, fondé sur le langage XML lui-même. Le «système descriptif de ressources» RDF (*Resource Description Framework*), publié en février 1999, est l'équivalent pour les données Internet des catalogues pour les livres de bibliothèque. Sur le réseau, les métadonnées



2. RÉDIGÉ DANS LE LANGAGE XML, un document qui contient les programmes de cinéma d'une ville peut être visualisé sur des machines différentes. Des «feuilles de style» filtrent, réarrangent et affichent l'information : elles engendrent des documents comprenant des graphiques pour les ordinateurs de bureau, des fichiers composés exclusivement de texte pour les agendas de poche ou des messages vocaux pour les téléphones.