Les archives d'un réseau actif

BREWSTER KAHLE

L'archivage des documents créés et échangés sur le réseau Internet pourrait être historiquement, commercialement et politiquement important.

Au Ier siècle avant notre ère, les manuscrits de la bibliothèque d'Alexandrie ont disparu dans un incendie ; les premiers livres imprimés sont tombés en lambeaux ; une grande partie des tout premiers films ont été recyclés, parce qu'ils contenaient de l'argent... Malheureusement, l'histoire pourrait se répéter avec le réseau Internet.

Personne n'a tenté d'archiver intégralement les textes et les images contenus dans les documents qui apparaissent sur le Web. Comment, alors, éviter la perte de précieuses informations historiques, culturelles et scientifiques?

Avec la diminution constante des coûts du stockage numérique, un petit groupe de techniciens, dotés d'ordinateurs personnels et de dispositifs de stockage de données, pourrait archiver en permanence les documents diffusés sur le réseau mondial. Avec quelques collègues, nous avons commencé cette tâche colossale voilà un an, dans le cadre du projet *Internet Archive*.

Lorsque ce numéro de *Pour la Science* sera en kiosque, nous aurons enregistré quasiment toutes les parties du réseau qui sont librement et techniquement accessibles. Cet ensemble de données, soit 2 000 milliards d'octets (deux téra-octets), comprendra du texte, des images et des sons (la Bibliothèque nationale de France contient l'équivalent de 20 téra-octets de texte). Dans les mois à venir, nos ordinateurs et nos supports de stockage enregistreront d'autres secteurs d'Internet, dont les informations du *Gopher* (un système d'information antérieur au Web) et les forums de discussions d'*Usenet*. Les enregistrements faits à ce jour se sont déjà révélés très utiles pour les historiens. Dans l'avenir, ils pourraient constituer le fonds d'une bibliothèque minutieusement indexée... et consultable par le réseau.

Dix personnes qui participent au projet travaillent dans une base militaire reconvertie, à San Francisco. L'ordinateur collecteur avec lequel nous recueillons les informations est situé au Centre de calcul de l'Université, de San Diego.

Nos programmes «sillonnent» l'Internet, recueillant des documents nommés «pages», site après site. Lorsqu'une page est enregistrée, ils recherchent des références croisées, des «liens» avec d'autres pages. Ils utilisent ces liens – des adresses insérées dans la page d'un document – pour se déplacer vers les autres pages, refont alors une série de copies et cherchent d'autres liens vers de nouvelles pages. Naturellement, ils vérifient que les adresses qu'ils se proposent d'examiner n'ont pas encore été visitées à l'aide des identificateurs de sites nommés «localisateurs unifiés de ressources», ou URL. Les programmes tels qu'*AltaVista*, de la Société *Digital Equipment*, emploient eux aussi un logiciel collecteur pour indexer les sites Web.

Notre projet est possible parce que le coût de stockage de données est en constante diminution. Le prix d'un giga-octet (un milliard d'octets) d'espace sur un disque dur revient à 1 000 francs, et le stockage sur bande magnétique revient à 100 francs par giga-octet. Pour stocker les données que les utilisateurs des archives risquent de rechercher fréquemment, nous avons choisi le stockage sur disque dur, tandis que pour les données moins utilisées, nous avons opté pour un dispositif qui charge et lit les bandes magnétiques automatiquement : en moyenne, un lecteur de disques met 15 millisecondes pour accéder aux données, contre quatre minutes pour un lecteur de bandes. L'information fréquemment consultée pourrait être composée des documents historiques ou d'un ensemble de pages qui ne sont plus en service.

Nous comptons mettre à jour nos archives tous les deux ou trois mois. Le premier archivage complet a duré environ un an, mais, lors de nos prochaines recherches, nous ne mettrons à jour que les informations modifiées depuis notre précédente inspection.

Nous n'aurons pas tous les textes, images, films et autres données du réseau, car les programmes collecteurs tels que le nôtre n'ont pas accès à des centaines de milliers de sites : les éditeurs limitent l'accès à certaines données ou stockent des documents dans un format inaccessible. Toutefois, même si notre archivage est incomplet, il donne un bon aperçu du réseau à l'époque de l'archivage.

Lorsque nous aurons stocké la partie d'Internet accessible au public, quels seront les services offerts par nos archives? Tout d'abord, nous fournirons des documents qui ne sont plus disponibles chez les éditeurs, ce qui est un service important si le standard d'hypertexte du réseau s'impose dans le monde savant. Nos archives pourraient également se révéler très utiles pour les entreprises. En outre, les données collectées serviraient de «copies d'archives» pour les services publics ou pour d'autres institutions possédant des documents accessibles au public. Ainsi, progressivement, nous disposerons d'une bibliothèque numérique.

Retrouver les liens manquants

Des historiens utilisent déjà ces archives. David Allison, de la *Smithsonian Institution*, y a trouvé des informations pour réaliser une exposition sur l'emploi des sites Web lors des élections présiden-