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-

NMAH, We the People : Winning the Vote exhibition

INTERNET ARCHIVE **a fourni au Musée d'histoire américaine du *Smithsonian* un ensemble de pages Web consacrées à l'élection présidentielle américaine de 1996. L'ordinateur qui accède aux pages a fait partie d'une exposition sur les campagnes présidentielles.**

tielles : les expositions du passé présentaient des spots télévisés de campagnes électorales ; elles s'enrichissent aujourd'hui de pages Web.

La constitution d'archives pose plusieurs problèmes allant du respect de la vie privée à celui des droits d'auteur. Que se passerait-il si une étudiante créait une page montrant des photographies de son petit ami du moment et si, plus tard, elle voulait «déchirer» ces photos, alors qu'elles sont toujours présentes dans les archives? Un homme politique pourra t-il effacer des données publiées durant ses années d'études? Plus généralement, l'accès public aux informations collectées est-il un abus des clauses d'«usage loyal» définies par la loi sur la protection des auteurs? (voir *Les droits d'auteur des œuvres numériques*, par Ann Okerson, *Pour la Science*, septembre 1996).

Pour résoudre ces difficiles problèmes, nous autorisons les auteurs à exclure leurs œuvres de nos archives. Nous envisageons également de communiquer aux chercheurs non pas des documents individuels, mais uniquement un recensement des grands thèmes archivés : on pourra, par exemple, compter le nombre de références aux pachydermes sur le Web, mais non regarder une page spécifique concernant les éléphants. Nous espérons que ces mesures suffiront à dissiper dans un premier temps les inquiétudes concernant le respect de la vie privée et des droits de la propriété intellectuelle. En testant sur l'Internet des concepts tels que celui d'usage loyal, nos résolutions des problèmes rencontrés au fil du temps lors de la mise en place d'*Internet Archive* contribueront peut-être à faire aboutir les débats politiques sur le respect de la vie privée et sur la propriété intellectuelle.

Avec *Internet Archive*, nous cherchons également à résoudre les problèmes de pérennisation de l'information présente sur le réseau. À Washington, la Commission «préservation et accès» cherche à éviter les pertes de données résultant des changements de standard du stockage numérique. Avec d'autres équipes, nous cherchons des critères techniques qui permettraient d'attribuer un nom d'identification unique (URN, «nom de ressource unité») aux documents numériques. À la différence d'un URL, qui caractérise son emplacement sur le réseau, un URN caractérise le contenu d'un document. L'attribution d'un URN à un document permettrait de le retrouver après disparition du site qui le contenait : on estime à une quarantaine de jours la durée de vie moyenne d'un URL. Un URN pourrait localiser d'autres URL donnant encore accès au document désiré.

D'autres projets d'archivage s'intéressent à des parties spécialisées du réseau : *DejaNews* enregistre les messages diffusés sur les forums *Usenet*, et *InReference* archive les listes de discussions par courrier électronique. Ces deux services sont financés par la publicité, ce qui pourrait également être un jour le cas d'*Internet Archive*. Aujourd'hui encore, nous finançons notre projet avec l'argent provenant de la vente d'une société de logiciels et de services pour l'Internet. De grandes entreprises d'informatique nous ont également donné du matériel.

De nombreuses années s'écouleront avant qu'une infrastructure ne préserve correctement les informations d'Internet et que les problèmes liés à la propriété intellectuelle ne soient résolus. Entre-temps, nous poursuivons notre archivage, car si trop de documents disparaissaient du réseau, nous perdrions toute trace de la naissance de ce nouveau média.

Brewster KAHLE a fondé *Internet Archive* en 1996. Il a créé, en 1989, le système de recherche de documents WAIS *(Wide Area Information Server)*.