ACTUALITES
à la une
international
europe
france
société
régions
carnet
horizons
analyses & forums
entreprises
communication
aujourd'hui
météo
sports
sciences
culture

version texte

le desk
les dépêches
les dossiers
les fiches pays
les thématiques
la check-list
les portfolios

CHAINES
aden
éducation
finances
forums
interactif
mots croisés / jeux

ANNONCES
emploi
immobilier

newsletters
vos préférences
aide et services
qui sommes-nous ?


édition électronique
éditions nomades
archives
les thématiques
abonnements

• LE MONDE | 05.04.02 | 10h04
• MIS A JOUR LE 05.04.02 | 12h19

Internet cherche à se préserver de l'amnésie

L'espérance de vie moyenne d'une page sur le Web se compte en semaines. Plusieurs initiatives ont vu le jour pour tenter de sauver de l'oubli cette formidable bibliothèque, toujours mouvante. La tâche s'annonce difficile.

"Les paroles s'envolent, les écrits restent." Média révolutionnaire, Internet a fait voler en éclats cet adage bimillénaire. Sur la Toile, l'espérance de vie moyenne d'une page va de quelques jours à quelques semaines. Passé un certain délai, cette page aura été modifiée, déplacée, ou simplement effacée. La masse monstrueuse d'informations portée par le réseau des réseaux est en perpétuelle mutation, constamment menacée de péremption. On reproche parallèlement au Web - marchand notamment - sa mémoire infaillible pour traquer l'internaute et prédire que, s'il a consommé ici, il paiera sans doute là. Mais ce *big brother* paradoxal est aussi un monstre d'oubli.

Brewster Kahle a été l'un des premiers à saisir l'ampleur de cette menace d'amnésie. Dès 1996, cet ancien du MIT, spécialiste des ordinateurs parallèles, a entrepris de stocker le Web à intervalles réguliers, sur son site archive.org. Son ambition : ressusciter la bibliothèque d'Alexandrie, qui avant ses incendies, contenait toutes les connaissance accumulées dans l'Antiquité. Fin 2001, Brewster Kahle, rejoint dans son initiative par nombre d'entreprises et d'institutions, a ouvert à tous les dix milliards de pages de ses archives. L'interface de son site, baptisée Wayback machine, permet de remonter le temps. Il suffit d'entrer l'adresse d'un site pour obtenir une liste de liens renvoyant aux précédentes versions engrangées depuis 1996 par les robots d'Internet Archive.

La masse de données à portée de souris est monstrueuse : plus de 100 téraoctets, l'équivalent de 100 millions de livres. Soit cinq fois plus que la bibliothèque du Congrès américain, associée au projet. La plus grande base de données jamais construite s'appuie, c'est la beauté de la chose numérique, sur 400 PC ordinaires qui n'ont coûté qu'un demi-million de dollars, là où la librairie du Congrès consomme annuellement 450 millions de dollars.

UN EXPLOIT TECHNIQUE

Chaque mois, les automates qui circulent pour engranger un nouvel instantané du Web "rapportent" 10 téraoctets ; l'équivalent de la moitié du contenu de la vénérable institution... L'exploit technique est remarquable. Mais passée la curiosité de voir combien certains sites familiers ont pu évoluer au fil des ans, l'accès aux données se révèle souvent ingrat : de nombreux sites souffrent d'une présentation bancale ; des photographies ou illustrations ont disparu ; les liens internes des pages ne fonctionnent pas toujours. Mais sur- tout, hormis pour certains sujets, comme les élections présidentielles américaines de 2000 ou les attentats du 11 septembre, l'interface ne dispose d'aucun moteur de recherche. Si bien qu'il est vain d'espérer trouver une information autrement que par hasard.

Internet Archive est encore loin d'être la mémoire universelle que ses

dimensions gargantuesques laissent supposer. Peut-être l'ambition de Brewster Kahle est-elle démesurée ? D'une part parce qu'on ne connaît pas la taille réelle de la Toile. Ensuite parce que la collecte automatique des sites se heurte à des obstacles techniques et juridiques innombrables.

La taille du Web, tout d'abord. Elle fait l'objet des estimations les plus diverses. Fin 2001, on comptait près de 9 millions de sites, dont 3,1 millions ouverts au public. Mais on dénombrait parallèlement 36 millions de noms de domaine (source *Netnames*, octobre 2001) déposés. Le nombre de pages visibles dépasse les deux milliards (Google effectue ses recherches sur 2,073 milliards de pages) et le Web s'enrichirait de un à sept millions de pages par jour. Mais le Web invisible, ou Web profond, auquel les automates qui parcourent la Toile de lien en lien pour le compte des moteurs de recherche n'ont pas accès, serait quatre à cinq cents fois plus vaste.

RÉTICENCES CROISSANTES

Ces chiffres, publiés en septembre 2001 par la société Bright Planet, qui propose une technologie capable de sonder ce Web profond, sont à prendre avec précaution. Mais ils illustrent l'une des difficultés auxquelles seront confrontés les futurs archivistes d'Internet : celui-ci devient de plus en plus dynamique, son contenu étant toujours plus personnalisé, modelable en fonction des demandes de chaque internaute. Sans compter qu'en combinant la ra- dio et la télévision, il s'apparente toujours plus à un média de flux. Espérer en prendre une "photographie" pertinente à un instant donné est illusoire.

La collecte automatique fondée sur l'utilisation des automates est loin d'être parfaite : les robots ne peuvent passer outre les multiples formulaires qui agrémentent un nombre croissant de pages d'accueil. Les sites payants ou à abonnement, même gratuit, restent hors de leur portée. Les gestionnaires des sites eux-mêmes peuvent les interdire d'accès (par une commande nommée robot txt). Cette procédure s'est multipliée à l'encontre des automates d'Internet Archive, après que le site a été ouvert au public : les éditeurs en ligne - notamment les journaux -, ont soudain pris conscience que leur contenus pouvaient être "aspirés" pour devenir accessibles gratuitement, alors qu'eux-mêmes peinent à commercialiser ce fonds de commerce.

Le coup de force de Brewster Kahle, qui a mis en ligne d'autorité tous les contenus ayant trait aux attentats du 11 septembre, a refroidi nombre d'éditeurs, qui n'avaient sans doute pas vraiment pris conscience de l'existence de cet Internet bis. La tradition anglosaxonne du *fair use*, l'"usage raisonnable", qui atténue les rudesses de l'usage du copyright et du droit d'auteur, a sans doute touché là ses limites. Un grand nombre de sites, parmi lesquels la NASA et de nombreux sites gouvernementaux, se sont "déréférencés" d'Internet Archive.

La Maison Blanche elle-même a disparu en novembre 2001 d'Internet Archive. Volonté de contrôler discours et agendas politiques, mais aussi peut-être d'empêcher des retours en arrière parfois éclairants ? Toujours est-il qu'Internet Archive pose de façon inédite la question de la maîtrise de l'information, ressource stratégique. L'autre question majeure lancée dès le début de l'initiative de Brewster Kahle - celle, juridique, du respect de la propriété intellectuelle - risque de trouver une réponse tout aussi pragmatique. Elle est d'ailleurs fournie par le site d'Internet Archive lui-même : ceux qui le veulent peuvent disparaître des mémoires de la Wayback Machine. Il suffit de le demander.

Hervé Morin

La mémoire des physiciens

Les physiciens, qui ont grandement contribué à l'essor d'Internet - Tim Berners-Lee, du CERN, est à l'origine des liens hypertexte -, ont utilisé très tôt cet outil pour échanger leurs connaissances. Dans le même temps, ils se sont

souciés de garder trace de leurs travaux. En octobre 1994, Paul Ginsparg, du Los Alamos National Lab, a ainsi mis à la disposition des chercheurs un serveur, baptisé "xxx", où ils pouvaient entreposer leurs documents et les mettre en ligne avant publication éventuelle dans des revues scientifiques. Cette initiative a rencontré un succès considérable. Des sites miroirs se sont installés dans le monde entier. En France, le CNRS a créé, en octobre 2000, un Centre pour la communication scientifique directe (CCSD) fondé sur ce modèle (http://ccsd.cnrs.fr/). Pour son directeur, Franck Laloë, du Laboratoire de physique de l'ENS, l'enjeu est de "préserver à long terme une information gratuite et universelle". Mais, prévient-il, "transmettre de génération en génération ces documents bruts ne sera pas évident."

• ARTICLE PARU DANS L'EDITION DU 06.04.02