Web en stock

Louise Merzeau

LOUISE MERZEAU

Web en stock

Cet article s'appuie sur un entretien avec Jean-Michel Rodes, directeur de l'Inathèque de France. Qu'il soit ici remercié pour les informations et réflexions précieuses dont il a bien voulu me faire part.

Dans une médiasphère dont le centre de gravité se déplace vers des systèmes eux-mêmes dépourvus de centre (réseaux de serveurs enchevêtrés) le paradigme de la transmission doit *intégrer le passage à la passation*. L'intérêt croissant des institutions patrimoniales envers des contenus jusqu'alors voués au dépôt sauvage ou à l'oubli témoigne en ce sens d'une prise de conscience qui a souvent fait défaut sous la vidéosphère. Faisant suite à la législation sur les archives audiovisuelles, le projet d'un Dépôt Légal d'Internet représente à ce titre une importante étape dans l'*accréditation culturelle de l'éphémère*. Archiver le web, c'est reconnaître que le capital symbolique à transmettre ne saurait désormais se couper des flux de données sans perdre une part essentielle de sa dynamique et de ses contenus. Mais c'est aussi exposer l'archive à une *contamination* par l'éphémère, qui tend à modifier la logique et la portée du geste patrimonial.

L'impensé technique de la rémanence

C'est devenu un lieu commun que de considérer le web comme l'archétype de la volatilité. Média de flux où l'information passerait sans sillage, Internet consacre pour ses détracteurs comme pour ses défenseurs le règne de l'éphémère. Comparé à l'imprimé, on ne peut que constater l'instabilité de ses contenus. Non seulement 70 % des pages web ont une durée de vie inférieure à quatre mois, mais la structure hypertextuelle des documents répercute cette précarité dans l'ensemble du réseau. Aucun site ne peut en effet garantir la pérennité de ses informations dès lors qu'il est lui-même relié à d'autres sites, donc exposé aux innombrables trous noirs des « liens cassés ». Or passent aujourd'hui dans le réseau des informations indispensables à l'appréhension de notre temps, qui n'ont plus nécessairement d'équivalent sur d'autres supports. Laisser le flux effacer ses traces à mesure qu'il les sécrète reviendrait donc à renoncer à des pans entiers de notre histoire. Sans parler du dispositif lui-même, qui, en tant qu'organe d'élaboration et de navigation dans le savoir, mérite d'être conservé au même titre que n'importe quelle machine à communiquer.

Cependant, la volatilité du web n'est peut-être elle-même qu'un leurre. La circulation des données numériques exige en effet des procédures d'inscription qui radicalisent l'économie des traces plus qu'elles ne la suspendent. « Découplant la vue logique de l'information de son implantation physique, la délocalisation des mémoires impose une forte redondance,

notamment des systèmes d'adressage et de vérification, [entraînant] une inflation des couches internes de gestion de l'information » [1]. C'est le paradoxe de cette logique informationnelle, qui prône l'éphémère en indexant la valeur des messages sur le temps, tout en favorisant un développement sans précédent des mémoires externes. Anticipation, indexation, normalisation, connexion et récupération : tels sont les rouages de cette nouvelle temporalité, où toute instruction, aussi temporaire soit-elle, s'enregistre nécessairement quelque part. De la standardisation des formats aux métadonnées, et des sites miroirs aux fichiers partagés, c'est tout un impensé machinique de la rémanence qu'il faut interroger, pour dépasser une opposition entre stock et flux, désormais inapte à décrire les nouvelles graphies du temps.
Comme le remarque Jean-Michel Rodes, le numérique serait en ce sens plutôt « à l'opposé de l'éphémère ». Si l'on n'est jamais certain de retrouver une page en pointant sur son URL, en raison de la caducité fréquente des liens (la fameuse « erreur 404 not found »), on ne peut davantage s'assurer de sa disparition. Le fait qu'une page ne soit plus signalée par les moteurs de recherche ne signifie pas qu'elle n'existe plus, et l'on peut d'ailleurs souvent retrouver sa trace en naviguant de lien en lien. Inversement, un fichier retiré du serveur pourra demeurer dans des caches et être encore référencé par des outils de recherche. Plutôt qu'un lieu d'amnésie, Internet est donc un *espace de réverbération*, où le signal ne disparaît que progressivement, par un phénomène « d'échos successifs qui vont en s'atténuant ».

Il faut par ailleurs éviter de confondre la réalité techno-économique du web avec l'image immatérielle ou lisse que cherche à en donner une idéologie de l'*immédiat* particulièrement active dans ce domaine. Exerçant un monopole de fait sur les discours qui ne versent pas dans la déploration nostalgique, celle-ci ne valorise souvent le bref et le fluide que pour mieux dissimuler la nature des stratégies industrielles en jeu. Les injonctions de l'actuel doivent alors être identifiées pour ce qu'elles sont : des arguments-écrans, masquant derrière la séduction des critères d'accessibilité ou de mobilité une programmation de l'obsolescence, relevant plus de la planification que d'un parti pris de l'éphémère. Par la promotion systématique de la vitesse et l'impératif de mise à jour, la compétition technologique et commerciale se donne en fait une façade présentable, tout en travaillant à disqualifier les médiations politiques, institutionnelles ou artistiques traditionnelles. Indexer la valeur d'une information sur sa fraîcheur et sa réactivité, c'est en effet court-circuiter les trajectoires hiérarchiques, nécessairement plus lentes, où se construit l'autorité.

Marché de l'éphémère, politique de l'archive

Le projet d'un Dépôt Légal d'Internet introduit en ce sens une double courbure dans l'économie temporelle du web. D'une part, il fait passer dans le champ de la *stratégie* des processus qui relevaient de la *tactique*, en convertissant un espace indéfini – celui des trajectoires imprévisibles et fugaces de l'information – en un lieu technocratiquement bâti, circonscrit et géré [2]. D'autre part, il offre une alternative aux stratégies de traçabilité actuellement en œuvre dans le réseau, en apportant la garantie de l'institution à une rétention jusqu'alors soumise aux seules lois de l'innovation technologique et de la concurrence. Législation, conservation et programmation n'auront donc pas seulement pour effet de relativiser la part éphémère du réseau. Elles permettront aussi de restaurer la portée *politique* de la mémoire comme de l'oubli. Paradoxalement, c'est peut-être là l'enjeu majeur d'un tel projet : protéger la société, non pas tant contre l'effacement de ses traces, que contre leur enregistrement systématique et hors de contrôle. Au terrorisme du temps présent, pourrait bien en effet répondre celui d'une traçabilité informatique « sauvage », tendant à déposséder

[1] Jean-Michel Rodes, « Les bouleversements de la mémoire au seuil du IIIe millénaire », *Médiamorphoses* n°1, janvier 2001.

[2] Sur cette distinction entre tactique et stratégie, voir Michel de Certeau, *Arts de faire*, 10/18, 1980.

le corps social de sa responsabilité mémorielle [3]. C'est pour faire contrepoids à la privatisation de la mémoire collective plus qu'aux effets d'une amnésie, qu'un archivage institutionnel est nécessaire.

Si elle fut pionnière en la matière, la fondation américaine Internet Archive [4] ne peut à ce titre présenter les mêmes garanties qu'un organisme public. La campagne de capture menée depuis 1996 par Brewster Kahle a bien permis de constituer une banque, librement accessible en ligne, de plus de 10 billions de pages aujourd'hui retirées du réseau. Mais cette « bibliothèque digitale », qui revendique des intentions culturelles, n'est pas à l'abri des pressions susceptibles de s'exercer sur toute entreprise relevant d'intérêts privés. Sous le prétexte du copyright, certains sites se sont ainsi vus retirés de la *wayback machine*, non pas à la demande de leurs auteurs, qui autorisaient et souhaitaient leur archivage, mais au terme d'intimidations émanant d'instances adverses ou concurrentes [5].
De leur côté, les outils de recherche pourraient compenser la volatilité d'Internet par la stabilité de leur index et la mise en cache des pages référencées, s'ils n'étaient exposés aux mêmes trafics d'influence. Les menaces de procès du *Monde interactif* à l'encontre de Google témoignent de ces interférences entre temps de l'information et de la marchandisation. En contestant au moteur de recherche le droit de conserver une copie des pages indexées (afin que les internautes puissent les consulter indépendamment de leur mise à jour), le journal entend monopoliser le contrôle de son obsolescence, pour mieux la rentabiliser [6]. À ces pratiques de discrimination par effacement, répondent symétriquement des opérations comme la vente au profit de Google des archives de Usenet [7]. Dans un cas comme dans l'autre, on voit le danger qu'il y aurait à déléguer la démarcation entre éphémère et perpétuité aux seules lois du marché.

Face à ces dérives, la France bénéficie d'une tradition vieille de cinq siècles, qui a conduit le législateur à définir une politique d'archive à chaque révolution médiatique. Fort de cette légitimité patrimoniale (réaffirmée en 1990 dans la Loi sur l'Information, puis dans les trois textes qui régissent aujourd'hui l'économie numérique), le projet d'archivage du web est avant tout un facteur de garantie, scientifique et juridique.
L'effacement aléatoire des pages et l'instabilité des liens n'affectent pas seulement la pérennité des contenus *en ligne*. Par un effet de ricochets, lié à la banalisation des usages d'Internet, cette précarité gagne l'ensemble de la production intellectuelle – qu'elle soit universitaire, éditoriale ou médiatique. De plus en plus nombreux sont en effet les livres, articles ou thèses introduisant des renvois à des URL, alors même que ces adresses ont toutes les chances d'être périmées dès parution… Le besoin d'une référence et d'une autorité publiques excède donc l'espace virtuel du web. Les contenus ayant partout tendance à s'affranchir des intermédiaires traditionnels, c'est le principe même du « recours à la source, fondement de toute démarche scientifique » [8], qui est en jeu. Avec le Dépôt légal, on gagne l'assurance de pouvoir consulter une page, avec ses accès d'origine et une datation, *dans l'après-coup d'une lecture différée*. Certes, une telle fixation altère la dynamique du web : l'éphémère n'entre au patrimoine qu'en se dénaturant, comme une fugacité artificiellement conservée *au deuxième degré*. Mais c'est à ce prix que la question de la maîtrise des sources –

[3] Le film de Steven Spielberg, *Minority report*, propose une passionnante illustration de cette dépossession des traces (voir L. Merzeau, « Gratteurs d'images », *Cahiers de médiologie* n°15, 2002).
[4] The Internet Archive : http://www.archive.org/index.html
[5] Ce fut notamment le cas pour des sites mettant en cause l'Église de scientologie, comme antisect.net ou xenu.net.
[6] Plus radicalement encore, le Groupement des Éditeurs en Ligne souhaiterait une réglementation interdisant qu'une URL puisse être mentionnée sans l'autorisation de l'éditeur.
[7] Cf. Emmanuel Hoog, « Internet a-t-il une mémoire ? », *Le Monde*, 16 août 2002.
[8] *Ibidem*.

fondamentale pour une société de l'information – cessera peut-être de passer au second plan…
Les hébergeurs et fournisseurs d'accès réunis au sein de l'AFA voient eux-mêmes dans le Dépôt Légal une certification contre les formes de piratage, de pistage et de pillage, qui ne pourra que renforcer leur propre autorité. Car derrière la question des sources, se pose celle des droits relatifs à l'anonymat des données comme des parcours de navigation. Notamment, faut-il conserver les pages personnelles au même titre que les autres contenus, ou les « abandonner » à l'éphémère en considérant que patrimonialisation rimerait ici avec indiscrétion ? L'institution a finalement jugé qu'elle ne pouvait écarter une telle richesse d'information, d'autant plus significative qu'elle joue un rôle moteur dans l'expansion du web. Surtout, si aucune instance publique ne les conservait dans des conditions légales, elles ne seraient pas effacées, mais livrées aux stratégies marchandes. Déjà, des sociétés se spécialisent dans le pistage des internautes, pour constituer des bases de données destinées à être monnayées. Contre l'atteinte à la vie privée ou l'intégrité des personnes, comme contre l'effacement arbitraire des informations, seule une mémoire publique offre la garantie d'un contrôle démocratique.

Quand le stock s'ouvre au flux

Une fois posée sa légitimité, le Dépôt Légal doit être envisagé sous l'angle de sa faisabilité. Comment stocker du flux, circonscrire un espace en croissance exponentielle, collectionner des objets non finis et pérenniser des dispositifs évolutifs ?
Si le réseau accroît toujours son périmètre, la question des volumes pose moins de difficultés qu'on pourrait le croire. D'abord parce que cette poussée se stabilise (on est passé de 82% en 1996 à 18% en 2001). Ensuite parce qu'en termes d'encombrement et de dénombrement, on est encore loin des quantités que doit traiter l'archivage des documents papier aussi bien qu'analogiques. Plus délicate est la question de la déterritorialité. Rebondissant de machine en machine, les documents à collecter ne se rapportent plus à un support d'inscription ou un lieu de publication, mais à des chemins de navigation où l'unité de sens se redéfinit à chaque bifurcation. Sans même considérer la vitesse d'actualisation des contenus électroniques, il faut désormais prendre en compte cet *inachèvement* propre à toute hypertextualité. Enfin, l'innovation technologique pose à l'archive un dernier défi : en même temps que la conservation des contenus, l'institution doit anticiper le vieillissement des programmes, appareils et standards d'encodage, ainsi que leur compatibilité. Autant dire qu'en passant dans la durée patrimoniale, le web change de nature tout en modifiant lui-même la grammaire archivale.

La sélection est le premier principe documentaire à repenser. Si le cadre légal circonscrit, comme pour les autres médias, une frontière nationale, on sait qu'Internet rend délicate une telle délimitation. Le nom de domaine en *.fr* ne permet d'identifier qu'une partie du web français [9]. La collaboration des sociétés délivrant des noms de domaine en *.com, .net*… est donc nécessaire pour dessiner les contours du territoire à contrôler. Mais la localisation du siège social des éditeurs ne dispensera pas l'archive publique d'assimiler la porosité de ses propres marges, tant Internet résiste à toute clôture. Non pas qu'il soit étranger à toute territorialité (là plus qu'ailleurs, la mondialisation se traduit par des luttes visant à s'assurer une dominance économique, sécuritaire, juridique, etc.). Mais la mémoire du réseau doit se penser elle-même comme *réseau de mémoires*. Au-delà des politiques nationales, une articulation des institutions patrimoniales sera nécessaire au plan mondial (pour normaliser par exemple l'écriture des URL chronologiques), afin que la carte de l'archive reflète au plus près la complexité du territoire archivé.

[9] Environ 150 000 sites selon l'AFNIC, l'Association Française pour le Nommage Internet en Coopération.

Le mode de collecte est le deuxième indicateur de cette contamination du stock par le flux. Contrairement aux pays ayant choisi le dépôt volontaire ou la collecte manuelle [10], la France a opté pour une méthode d'aspiration automatique des sites. Il faut s'imaginer des centaines d'agents logiciels autonomes travaillant sur des grappes de PC et dialoguant avec une base de données gérant le plan global de captation. 90% des sites, écrits en html ou relativement statiques, pourront ainsi être automatiquement récupérés (certaines machines étant dédiées aux sites les plus « imposants », les autres traitant plusieurs centaines d'adresses). Restent 10% de sites dynamiques ou d'accès protégé, dont la capture nécessitera des procédures d'échantillonnage ou d'archivage manuel. Les sites en mode *streaming* [11] seront quant à eux captés en continu. Pour l'INA, il est envisageable de se rapprocher d'une capture quotidienne du web, avec un dispositif technique qui modulerait cette fréquence selon les types de site et les itérations [12]. Dans l'ensemble, même si la loi prévoit un principe de sélection, c'est donc la logique de l'aspiration qui l'emporte, l'exhaustivité étant limitée par des critères d'économie plus que de pertinence.

Étape suivante : la définition de l'unité documentaire. N'étant plus assigné à un support durable, le document se segmente en éléments plus ou moins autonomes (images, boutons, bandeaux, textes…), que l'on doit traiter isolément. C'est la quantité de ces objets à rapatrier (plusieurs milliards) plus que le nombre d'octets qui entraîne un formidable changement d'échelle de l'archive. La solution envisagée consiste à doter chaque objet d'une signature définitive, pour éviter de le recalculer à chacune de ses occurrences, et limiter ainsi la redondance des données. Au stockage de ces unités s'ajoute le recensement des liens internes et externes au site, par lequel sera préservée la dimension opératoire de l'ensemble. Chaque parcours de navigation recevra alors une étiquette chronologique, attestant des états dans le temps d'un système toujours en équilibre – le stock dessinant ainsi une sorte de *cartographie horaire* du flux.

Pour finir, l'archive devra s'assurer de sa propre pérennité, en opérant de régulières migrations des supports de stockage, des langages et des formats. Fragilisée par l'innovation technologique, la conservation doit en effet contrebalancer ici la vitalité électronique par la production de standards ou d'algorithmes de conversion, tout en s'ouvrant elle-même au principe de l'évolutivité.

Une culture distribuée

À terme, le web archivé sera mis à disposition des citoyens, dans des condition de consultation analogues à celles des archives audiovisuelles, c'est-à-dire dans des lieux accrédités tels que bibliothèques ou universités. Si elle pouvait sembler aller de soi, la mise en ligne d'un tel stock a en effet été écartée, parce qu'elle positionnerait l'institution dans une situation de concurrence absurde avec les moteurs de recherche. Surtout elle tendrait à confondre la carte et le territoire, en accentuant les effets de mise en abîme qui affectent l'archive dès lors qu'elle s'ouvre à l'éphémère. On a préféré miser sur le temps, qui creusera progressivement le nécessaire écart entre le flux et son double archivé.

Le dernier retournement que favorisera sans doute un tel projet concerne l'acculturation. On peut s'attendre à ce que le temps court gagne ici ses lettres de noblesse aux yeux d'une raison

[10] C'est notamment le cas de l'Australie et du Canada.

[11] Les sites dynamiques comportent des pages où les informations sont automatiquement changées en fonction d'une base de données ou d'éléments provenant de l'utilisateur (c'est notamment le cas des URL en .asp, .cfm, .cgi ou .shtml). Le *streaming* permet quant à lui de consulter des documents de flux (audio ou vidéo) en continu, sans attendre le téléchargement complet des données sur le disque servant à la consultation.

[12] À titre de comparaison, la captation automatique du web mise en œuvre par la Bibliothèque royale de Suède dans le cadre du projet Kulturar W3 n'est alimentée que deux à quatre fois par an.

qui ne se pense encore que dans la durée. La perspective patrimoniale incitera peut-être les chercheurs à construire autour du web un espace de réflexion qui, selon Jean-Michel Rodes, tarde encore à se développer dans les sciences humaines comme chez les décideurs politiques.

Mais plus fondamentalement, c'est la fonction archivale elle-même qui sera probablement réévaluée. L'usage d'Internet contribue en effet à distribuer de plus en plus largement des compétences documentaires jadis confinées dans certains corps de métier. Les savoir-faire relatifs à la formulation de requêtes, au contournement du bruit, à la granularité de l'information ou au référencement (comme à leur subversion) remontent ainsi *en amont de la documentation*. Contrairement aux objets audiovisuels, les contenus électroniques comme les parcours à conserver sont déjà structurés dans une perspective d'accès et de traitement. Inversement, les volumes à rapatrier contraindront de plus en plus l'archive publique à se limiter à des indexations automatiques, incapables d'atteindre les couches sémantiques fines que le traitement manuel des documents textuels ou analogiques atteignait [13].
C'est l'ultime pirouette de l'éphémère, que d'introduire ainsi dans les pratiques du bref un réflexe d'archivage et d'anticipation, par où le stock irrigue le flux, avant même qu'il fasse l'objet d'une politique de mémoire.

[13] Les recherches actuelles sur le *web sémantique* incitent à penser que c'est le flux lui-même qui assurera cette fonction d'indexation fine, en complétant le contenu informel du web actuel par de la connaissance formalisée.