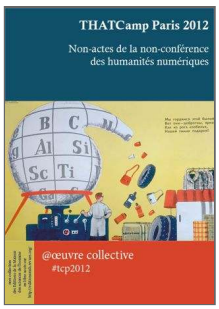

THATCamp Paris 2012
Non-actes de la non-conférence des humanités numériques

Éditions de la Maison des sciences de l'homme

Sommes-nous en train de perdre la mémoire ? Mémoire et archivage du web

Proposé par Frédéric Clavert et Clément Oury

Collectif

DOI : 10.4000/books.editionsmsh.309
Éditeur : Éditions de la Maison des sciences de l'homme
Lieu d'édition : Paris
Année d'édition : 2012
Date de mise en ligne : 1 octobre 2012
Collection : La Non-Collection
ISBN électronique : 9782735115273

Sommes-nous en train de perdre la mémoire ? Mémoire et archivage du web

Proposé par Frédéric Clavert et Clément Oury

Collectif

Introduction

1 En 2003, dans son article « *Scarcity or Abundance ? Preserving the Past in a Digital Era* »[1], Roy Rosenzweig attirait l'attention sur les enjeux de la préservation des traces numériques du passé. Malgré l'abondance des sources désormais disponibles sur le web, il existe un risque réel de perdurer dans une rareté de l'information qui a toujours été, et de ne pouvoir, à l'avenir, que construire une histoire fragmentaire de notre époque. Insistant sur l'urgence de trouver les moyens de pérenniser les sources numériques primaires, il montrait les limites de certaines techniques, notamment l'impression des sources numériques (l'*Auswärtiges Amt* à Berlin procède de cette manière pour les courriers électroniques), la conservation d'un lecteur pour les supports obsolètes (pratiqué par l'Institut national de l'audiovisuel en France, INA), la migration régulière d'un support ou d'un logiciel à l'autre ou la virtualisation. Pour Rosenzweig, une des solutions est la mobilisation commune des historiens et archivistes, comme celle qui a permis, en 1934, la création de la *National Achives and Records Administration* aux États-Unis avant l'éloignement de leurs organisations professionnelles respectives[2]. Ce dernier point rappelle l'une des conclusions de THATCamp Paris 2010 et l'une des particularités des *Digital Humanities* : la nécessité de l'interdisciplinarité et du travail en commun de multiples corps de métier.

2 L'article de Roy Rosenzweig date de 2003. Depuis, le web a évolué vers plus de « Web 2.0 » et, notamment, vers les réseaux sociaux comme Twitter ou Facebook, pour nommer les plus connus. La question de la préservation des données publiées sur les réseaux sociaux

se pose avec autant d'acuité que celle du site satirique *Bert is evil* pris en exemple par Rosenzweig, où il souligne que la préservation de ces données pose deux questions principales. Doit-elle être laissée à des organismes privés – comme Internet Archive et surtout les entreprises possèdant les réseaux sociaux dont l'intérêt est, avant tout, l'exploitation commerciale de ces réseaux et, certainement pas, l'archivage à des fins d'analyse historique ? Qui peut accéder à ces données et quelle législation va encadrer cet accès, entre la loi des trente ans s'appliquant aux archives publiques et un « droit à l'oubli » en émergence[3] ?

3 Un article récent[4] démontre que les ressources partagées sur les réseaux sociaux disparaissent au rythme de 0,02 % par jour. Depuis la fin des révolutions libyenne, égyptienne et tunisienne du printemps 2011, environ 30 % des sites partagés *via* les réseaux sociaux ont tout bonnement disparu. En clair, la situation décrite par R. Rosenzweig ne s'est que peu améliorée. De plus en plus de bibliothèques nationales archivent des portions du web, au titre du dépôt légal. Cet archivage vient s'ajouter à celui d'entreprises privées comme Internet Archive et sa *wayback machine* toutefois quelque peu négligée ces derniers temps. Des logiciels pour la préservation de collections numériques, appelés *repository* (Fedora, DSpace, Alfresco[5]) permettent, bien utilisés, de gérer des millions de documents de manière pérenne (si le financement est bien entendu aussi pérenne). Enfin, des identifiants uniques ou permanents apparaissent : le DOI (*Digital Object Identifier*), devenu une norme ISO en 2012, implémente le système Handle, les adresses web permanentes de type purl, etc.[6].

4 Dans une perspective plus historique, la question de la gestion des archives a toujours été double. D'un côté, se pose celle de la sélection des archives, un problème important dans le domaine numérique à l'ère de « l'infobésité », de l'autre, celle de leur préservation. La perte des archives n'est pas un phénomène nouveau. En Égypte, par exemple, nous disposons des papyrii qui ont été préservés dans les zones sèches et désertiques. Ceux qui ont été conservés dans des endroits trop proches du Nil, en zone humide, n'existent plus. Les guerres ont toujours fait disparaître des archives, soit de manière directe à cause des combats, soit de manière indirecte en raison de leurs conséquences politiques. Certaines disparitions furent temporaires. Ainsi, la division est/ouest pendant la Guerre froide a bloqué pour les historiens de l'Ouest l'accès à de nombreuses archives du régime nazi qui avaient été emportées par les Soviétiques. Mais, la multiplication des sources numériques donne une dimension nouvelle par sa masse à ces problèmes classiques pour les archivistes et les historiens.

5 À travers plusieurs questions principales, cet atelier se concentre sur la spécificité de cette problématique de l'archivage et de la perte à l'ère du numérique, notamment à propos de la mémoire du Web. Peut-on mesurer l'ampleur des pertes d'archives, et les comparer à celles que nous avons déjà connues auparavant ? Quelles solutions de préservation et d'archivage émergent ? Si nous devons sélectionner les archives numériques à conserver, quels critères vont être adoptés ? Vont-ils différer de ceux utilisés pour la sélection des archives de l'ère « papier » ? Pouvons-nous faire le choix de l'abondance, c'est-à-dire de tout garder ? On rejoint là un questionnement traditionnel, mais qui a sans doute une résonance spécifique dans le cas des sources numériques. Enfin, un atelier animé par G. Poupault avait déjà abordé la question lors du THATCamp de Paris en 2010, soulignant qu'il fallait « archiver le temps » selon l'expression de Frédéric Clavert, puisque le document numérique, fluide, change de forme et de contenu. Il

faudrait donc conserver non seulement le document, mais aussi tout son historique. Pouvons-nous le faire ?

Archiver le Web : la politique de la BNF

6 Clément Oury présente tout d'abord le dépôt légal du Web dont la BNF est chargée depuis la loi relative au droit d'auteur et aux droits voisins dans la société de l'information (DADVSI) de 2006, parallèlement à l' INA qui s'occupe de collecter les sites de la communication audiovisuelle[7]. La loi étend le dépôt légal, dont l'origine remonte à 1537 sous François I[er] et qui a d'abord été appliquée aux livres publiés en France, à tous les « signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique[8]. » La radio, la télévision et les logiciels (dont les jeux vidéo) étaient, de leur côté, soumis au dépôt légal, depuis 1992, et le besoin d'archiver le nouveau support qu'est le Web s'est fait sentir à partir du début des années 2000. La loi permet, en outre, l'archivage des sites sans qu'une autorisation soit à demander aux personnes qui éditent et produisent des sites sur le domaine national français. En contrepartie, la consultation des archives du Web s'effectue uniquement sur place.

7 Ce cadre juridique nouveau est très important : il signifie que les sources Web sont considérées comme des contenus patrimoniaux au même titre que les autres documents collectés par la BNF. Du point de vue de cette institution publique nationale, l'objectif est donc d'organiser la mémoire du Web en fonction de son savoir-faire et de sa mission patrimoniale traditionnelle. Mais, pour y parvenir, elle a aussi besoin aussi des regards externes et, notamment, de ceux des chercheurs qui sont les principaux utilisateurs de ces collections.

8 Cette mission est cependant difficile à mener sur un terrain aussi vaste et mouvant que le Web. D'une part, si le dépôt légal s'applique en théorie à ce qui est diffusé sur le plan national, cela n'a pas grand sens de parler de territorialité sur Internet. En pratique, il s'agit donc de collecter des sites produits et édités sur le « Web français » (le domaine national français en « .fr » et les sites des départements et des territoires d'outre-mer), ou les sites de personnes physiques ou morales domiciliées en France (sites en « .org », « .net », « .com », etc.). Ce n'est pas la langue qui définit le dépôt légal. Ainsi, un site s'exprimant dans une autre langue mais hébergé en « .fr » sera également concerné par l'archivage du Web français. D'autre part, le dépôt légal est censé être exhaustif. Mais, archiver le Web de façon exhaustive est bien sûr un objectif impossible. La BNF remplace donc l'exhaustivité par la « représentativité », notion qui suscite beaucoup de discussions lors de l'atelier.

9 Enfin, une autre distinction cruciale est apparue dans les échanges sur la définition du périmètre à archiver. Comment distinguer entre la nature privée et publique d'une source Web ? En effet, le dépôt légal s'applique à ce qui est publié, non aux échanges à caractère privé. Du point de vue de la BNF, un blog est une publication et non une conversation. Il est donc concerné par le dépôt légal. En revanche, un forum ne l'est pas et donc n'est pas conservé. Ceci peut certainement se discuter. Une correspondance privée, à l'ère du numérique, ne sera pas non plus archivée par la BNF et il faudra recourir à des services en ligne pour récupérer des boîtes mails. De même, les espaces privés des réseaux sociaux ou des sites intranet sont exclus. *A priori* la limite entre le public et le privé est posée par

l'internaute lui-même. Mais il est vrai que ces notions sont très floues quand elles s'appliquent à Internet, et sont en voie de redéfinition.

> Médiévizmes @medieviz
>
> @inactinique #tcp2012 on expliquera sur le pad pourquoi ne pas conserver un forum ?

10 Une contrainte forte influence ces choix et ces définitions : la méthodologie et les outils utilisés pour la collecte. Concrètement, celle-là est effectuée par des robots qui « moissonnent » le Web français, c'est-à-dire 2 millions de sites en « .fr », avec une profondeur de capture moyenne. Tout ce qui est collecté est mis sur un serveur. Certains sites sont « photographiés » une fois par jour si leur importance patrimoniale le justifie. C'est le cas, par exemple, des sites des grands quotidiens d'actualité. La discussion porte aussi sur d'autres types de collectes plus ciblées et liées à un grand événement ou à un thème précis. Dans ce cas, la moisson peut aller plus en profondeur. Les campagnes électorales en ligne, régionales, nationales ou européennes, sont ainsi préservées par la BNF. L'an dernier une collecte spécifique a été effectuée dans l'urgence sur les révolutions du « printemps arabe » entre décembre 2010 et mars 2011.

> Jean-Pierre Masse @jpmasse
>
> #tcp2012 la BNF et l'INA ont "des robots qui se promènent" et qui explorent le web

11 L'exemple est discuté assez longuement car il pose bien le problème de la constitution d'une archive du Web. Les sites sont majoritairement en français et non en arabe. Du point de vue de l'historien, il s'agit d'une source européano-centrée qui introduit un biais sur les événements. Clément Oury est conscient des limites du corpus, mais souligne qu'il faut bien avoir en tête les conditions matérielles de collecte du côté des archivistes. La BnF, tout comme les gouvernements, a été prise au dépourvu et a organisé une collecte dans l'urgence en ayant conscience de la fragilité de ces sites. Le personnel étant majoritairement francophone, il s'est en effet attaché en priorité à des sources en français. Cependant, le service des collections orientales a pu leur indiquer des sites en arabe qui ont ainsi été moissonnés. Une coopération internationale a pu être construite pour ajouter d'autres sources et renforcer les contenus de cette collection Web[9]. Sur Internet comme pour les archives non numériques, souligne-t-il, ce qui est important est d'expliquer clairement comment la collection a été constituée. Les historiens peuvent ensuite prendre en compte ces biais dans leur analyse des sources.

Des problèmes classiques qui se posent de façon plus aiguë

12 Il est intéressant de constater que les discussions de l'atelier sur la mémoire du Web abordent de grandes questions classiques sur la constitution des archives et le rapport des historiens à leurs sources. C'est le cas, tout d'abord, de la distinction entre privé et public. L'hébergeur d'un forum qui n'est pas pris en compte par le dépôt légal du Web va décider s'il archive ou non son site en recourant à des services privés. L'archivage des réseaux sociaux est avant tout constitué par les sociétés privées qui les développent. Or

les entreprises privées ont toujours eu une perspective différente de celles des pouvoirs publics dans la constitution de leurs archives. Elles gardent ce qui leur est utile, et non ce qui va être utile aux historiens. C'est le cas de BNP-Paribas, par exemple, qui voulait récemment jeter toutes ses archives liées à l'Indochine (puisque la Banque d'Indochine faisait partie d'une partie de son histoire sans intérêt pour la banque d'aujourd'hui). De la même façon, risquons-nous un jour de devoir aller voir chaque hébergeur (par exemple Twitter, qui ferme son API) pour accéder à des archives qui auront le biais de l'entreprise privée ? La nature numérique ou non de ces archives ne change pas fondamentalement le problème.

13 Sur ce plan, les collections Web de la BnF ont les limites de tout dépôt légal. Certains sites en sont exclus et, même pour ceux qui sont concernés par la loi, il ne s'agit pas d'une solution de pérennisation en tant que telle. Certes, si un site disparaît par accident et qu'il a été collecté, la BnF peut lui fournir une copie de secours, à charge pour lui de remettre en ligne le contenu. Mais ce n'est pas la fonction principale de l'archive du Web de l'institution. Des services commerciaux d'archivage et de sauvegarde spécifiques existent pour ce type de besoins. Enfin, certains participants à l'atelier évoquent les risques de « privatisation », ou du moins le caractère limité de l'accès à ces collections numériques puisqu'il faut aller sur place pour les consulter, en vertu de la loi DADVSI[10]. Cependant, c'est une contrepartie à la possibilité qui est donnée à la BNF de collecter les sites sans demander d'autorisation préalable à leurs auteurs – une exception à la règle habituelle –, ce qui limiterait beaucoup le nombre de sites effectivement collectés, et donc la qualité de l'archive finale[11].

14 La sélection de sources au moment de la constitution de l'archive est un autre problème récurrent. La position de la BNF présentée au début de l'atelier est de privilégier la « représentativité » des sources et non une impossible exhaustivité. Clément Oury souligne aussi que la perspective patrimoniale de l'institution est « d'anticiper sur les besoins du chercheur dans 500 ans ». Ces expressions, bien sûr, font réagir les participants. La représentativité est vue avant tout par la BNF comme une question de fréquence de collecte. L'indexation est automatique et se fait sur l'ensemble des sites pour les collectes larges. Il y a intervention humaine pour la sélection des sites de base utilisés pour les collectes thématiques ciblées, par exemple le Web politique en 2012. Dans ce cas, les sites sont fournis par une vingtaine de bibliothèques. En outre, les choix de sélection et d'indexation sont documentés et accessibles. La liste des URL des sites archivés depuis 2002 pendant les périodes électorales est ainsi accessible sur www.data.gouv.fr[12]. Un effort important est donc fait pour que le processus soit transparent.

> Frédéric Clavert @inactinique
>
> @medieviz échantillonnage? On sélectionne une partie des sources pour conservation ou on sélectionne un échantillon représentatif? #tcp2012

15 On sait bien, toutefois, que les sources considérées intéressantes à analyser par les historiens ne se limitent pas à ce que les administrations gouvernementales jugent ou ont jugé, à une certaine période, nécessaire ou pertinent de conserver. Les historiens des représentations ou de l'intime, entre autres, sont familiers de ces débats sur ce qui peut devenir source, même les éléments considérés comme insignifiants à un moment donné, ou du point de vue des institutions politiques. Ceci recoupe également une différence familière de perspective et de culture professionnelle entre les historiens, dont le premier

mouvement serait de tout conserver, et les archivistes qui sont conscients de la nécessité de faire des choix pour pouvoir matériellement conserver des collections d'archives dans de bonnes conditions. Bien sûr, « anticiper sur les besoins du chercheur dans 500 ans » peut sembler un objectif d'une « arrogance intellectuelle incroyable », comme le notent certains participants, mais c'est aussi le mandat attendu d'une institution patrimoniale nationale qui a pour devoir de faire des choix afin d'être opérationnelle. Cette formule est certainement propre à provoquer la réflexion et traduit bien le fait que, dans le domaine du numérique comme dans des fonctions patrimoniales plus classiques, on récupère ce qu'on choisit de garder. Il y a donc bien, à la base et dans toute politique publique d'archivage, un choix politique.

16 En fin de compte, sur ces deux points, les problèmes ne sont pas nouveaux mais ils acquièrent une dimension particulière en raison de la masse des données collectées sur le Web et de la fluidité des évolutions en cours. Cependant, d'autres enjeux débattus pendant l'atelier montrent qu'il existe une spécificité de l'archivage du Web qui nécessite de repenser la notion de préservation.

Repenser la notion de préservation

17 Un premier enjeu évoqué par les participants concerne l'accessibilité des données. Le stockage est une chose, mais la préservation des collections suppose aussi de mettre à disposition des outils pour les rendre utilisables par les chercheurs. Dans un contexte numérique, la question est particulièrement importante. Il faut fournir des outils de fouille des données, et la BNF réfléchit à une navigation dans les versions conservées, à une indexation de meilleure qualité, ou à des cartographies de sites avec Gephi, par exemple[13]. Mais ceci n'est pas encore disponible. Avec 17 milliards de fichiers, il est difficile d'avancer vite.

18 En outre, il faut assurer la lisibilité des archives numériques. Se pose donc le problème de l'évolution des formats et des logiciels, qui deviennent rapidement obsolètes. Il faut conserver à la fois la source informatique et la source logicielle. Pour les archives du Web, la migration vers un format nouveau fonctionne mal. On privilégie les logiques d'émulation pour lire le Web d'il y a cinq ans avec des outils reproduisant les technologies de l'époque. Avec un navigateur et quelques *plug-ins*, on peut lire 99 % des archives actuellement conservées. Cependant, toutes les procédures d'émulation ne sont pas encore bien établies. Il faut aussi faire des répertoires de formats, les documenter et stocker des exemples de formats et d'outils de lecture. La BNF travaille avec Microsoft sur les formats et les émulations, mais cela pose des problèmes juridiques considérables pour les logiciels propriétaires.

19 Même dans le cas où les formats ne sont pas devenus obsolètes, ils constituent un obstacle de taille dans l'accès aux archives du Web. En effet, les robots moissonneurs maîtrisent certains formats et pas d'autres[14]. Les données incluses dans des pages auxquelles ils ne peuvent accéder ou dans des formats illisibles pour eux ne seront donc pas conservées. Les chercheurs qui travaillent sur ces collections ont donc l'habitude d'utiliser des archives « à trous ».

20 La question technique et la masse même des données en jeu sur le Web posent de façon brutale la question des moyens et de leur affectation. Archiver le Web demande de gros équipements et des budgets conséquents qu'il faut prévoir bien en amont. Les sources

numériques sont pérennes à condition qu'un investissement suffisant soit prévu pour fournir une redondance du stockage en interne, et une réplication dans des sites secondaires. En pratique, il s'agit actuellement d'une annexe en banlieue, selon un système partagé avec le Centre informatique national de l'enseignement supérieur (CINES)[15]. Les projets ont été lancés récemment dans des collectes sophistiquées, où le budget nécessaire au stockage n'a pas été inclus, courent le risque de perdre leurs données (cela s'est déjà produit).

> Dépôt légal Web BnF @DLWebBnF
>
> La réplication est capitale ds le domaine de la conservation du numérique #tcp2012
>
> Jean-Pierre Masse @jpmasse
>
> #tcp2012 la BnF redonde en interne ;-)

21 De plus, les institutions comme la BNF doivent faire un second arbitrage : affecter plus de ressources à la collecte (plus d'archives disponibles plus tard aux chercheurs) ou à la valorisation des données actuellement conservées (ce qui permet un meilleur accès immédiat des chercheurs, mais au détriment de la quantité de données collectées, et donc des archives numériques disponibles plus tard). C'est un problème de financement et de choix, toujours difficile, entre le court et le long terme. La BNF a une position institutionnelle qui la pousse à accorder autant d'intérêt aux besoins du chercheur dans 500 ans qu'à celui d'aujourd'hui. On ne peut pas connaître les futurs utilisateurs des sources, mais on conserve pour eux. L'archivage du Web est donc nécessairement pris en charge par une institution capable mettre en place une organisation importante, ce qui introduit une distorsion inévitable pour les futurs historiens.

22 La BNF et le dépôt légal du Web ne peuvent couvrir qu'une partie des archives Internet. La plate-forme HAL conserve les productions scientifiques déposées dans le cadre des archives ouvertes, notamment la littérature grise[16]. Le CINES offre des solutions aux bibliothèques universitaires[17]. Les responsables des diverses institutions sont censés le faire pour leurs données. L'entrepôt SPAR (Système de préservation et d'archivage réparti) de la BNF propose également un service tiers-archivage[18]. Le TGE-Adonis est également un acteur important dans ce domaine. Son offre d'archivage, lancée en 2009, va être redéployée très prochainement, toujours avec le CINES. Des coopérations devraient être développées entre tous ces acteurs, mais il n'est pas toujours simple de les mettre en place, essentiellement pour des raisons institutionnelles (ils dépendent de ministères différents).

23 Enfin, même dans le cas de projets de taille bien plus modeste lancés par des chercheurs, la collecte et l'archivage de sources numériques nécessitent un soutien institutionnel fort. C'est le cas, par exemple, du projet *Bracero History Archive* qui collecte et présente des sources diverses sur l'histoire d'un programme bilatéral qui a, entre 1942 et 1964, organisé l'immigration de travailleurs mexicains aux États-Unis[19].

24 Au-delà de ces questions fondamentales de moyens et de taille institutionnelle de ces projets, la spécificité du web comme support, dominé par les flux de données et les fluctuations dans le temps, nous oblige tous, historiens comme archivistes, à repenser ce que nous entendons par « préservation ». À chaque lecture, l'archive doit être réactivée et recontextualisée. Il faut pouvoir documenter l'évolution de son contexte – à la fois celui

des choix de conservation, mais aussi celui des attentes et de *l'habitus* des lecteurs des sources. Ceci est particulièrement difficile à faire pour des sources du Web. Il manque des modèles de gestion des données permettant d'archiver aussi le temps du support de l'archive et le temps de sa réception. L'enjeu est donc de réinventer un modèle conceptuel autour de cette notion de préservation appliquée aux archives du Web.

Dépôt légal Web BnF @DLWebBnF

De la difficulté de pérenniser tout en continuant à donner accès et à documenter le contexte de consultation #tcp2012

Conclusion

25 Quatre points principaux se dégagent de l'atelier. Il est d'abord urgent de définir un modèle conceptuel de la préservation par l'accès (et de l'accès) aux archives du Web. C'est en fonction de ce modèle qu'il faudrait, dans un second temps, définir les missions à assurer au profit de l'utilisateur/lecteur. Le choix des moyens techniques nécessaires et des politiques de conservation devrait en découler. Enfin, les possibilités institutionnelles d'organisation de cet archivage devraient être développées dans un cadre de mutualisation au profit de tous les acteurs de l'enseignement supérieur et de la recherche. Aussi bien ces conclusions que le ton animé des discussions sont révélateurs du vif intérêt suscité par la problématique de la mémoire et de l'oubli sur le Web, mais aussi des divergences de perspectives qui subsistent entre archivistes et historiens.

26 Les suggestions de Rosenzweig en 2003 sont donc toujours aussi pertinentes. Le paradigme de la surabondance des sources introduit par l'ère numérique, même s'il a tout pour réjouir les historiens, constitue en réalité pour eux un redoutable défi. S'ils ne s'impliquent pas maintenant dans la réflexion sur la préservation des archives du Web et dans le lobbying en faveur de politiques d'archivage adéquates, en alliance avec les archivistes, ils se préparent des lendemains douloureux, paradoxalement dominés par la rareté plus que par l'abondance.

NOTES

1. *American Historical Review,* 108 (3), June 2003, pp. 735-762, reproduit dans le *Read/Write Book 2*, http://press.openedition.org/265.

2. Rosenzweig, *op. cit.,* pp. 759-761.

3. *Cf.* les débats en France en 2009-2010 autour de la « Charte du droit à l'oubli » proposée par le gouvernement, le nombre croissant de plaintes auprès de la CNIL en vertu de ce même droit à l'oubli sur Internet encore mal défini (voir le rapport d'activité 2011 de la CNIL, en ligne sur son site http://www.cnil.fr), et la proposition de règlement de la Commission européenne en janvier 2012 (COM(2012) 11 final, http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_fr.pdf).

4. H. M. Salah Eldeen et M. L. Nelson, « Losing My Revolution : How Many Resources Shared on Social Media Have Been Lost? », arXiv :1209.3026, septembre 2012.

5. Fedora Project (http://fedoraproject.org/), DSpace (http://www.dspace.org/) et Alfresco (http://www.alfresco.com/).

6. Voir le site de la DOI Foundation (http://www.doi.org/), celui du système Handle (http:// www.handle.net/) et de purl (http://purl.oclc.org/docs/index.html).

7. Loi n° 2006-961 du 1er août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information, http://www.legifrance.gouv.fr/affichTexte.do? cidTexte=JORFTEXT000000266350. Voir aussi le Code du patrimoine, Titre III, dépôt légal (partie législative) http://www.legifrance.gouv.fr/affichCodeArticle.do? cidTexte=LEGITEXT000006074236&idArticle=LEGIARTI000006845516 et la partie réglementaire (Article R132-23) http://legifrance.gouv.fr/affichCodeArticle.do? cidTexte=LEGITEXT000006074236&idArticle=LEGIARTI000025004800, qui date seulement de décembre 2011.

8. Loi n° 2006-961 du 1er août 2006, Titre IV, article 39.

9. Deux cents sites, blogs ou pages Facebook ont été sauvegardés en coopération avec Internet Archive, la Bibliothèque du Congrès, l'Université de Stanford, l'Université américaine du Caire, la British Library et la Bibliothèque d'Alexandrie. Voir http://blog.bnf.fr/lecteurs/ index.php/2012/07/02/la-revolution-du-jasmin-sur-la-toile/

10. Pour respecter le droit d'auteur, aucune diffusion en ligne de ces archives du Web n'est autorisée. On ne peut pas non plus copier librement les sources numériques consultées sur place. Il est possible d'imprimer les pages consultées, de prendre des notes ou de copier des extraits de texte, mais les captures d'écran ne sont pas autorisées.

11. Le dépôt légal du Web en Grande-Bretagne, par exemple, fonctionne sur un modèle différent qui permet des collectes bien moins riches. Le cadre législatif britannique protège les droits d'auteurs même dans le cas des publications numériques comme des sites Web. La British Library ne peut donc qu'encourager les propriétaires et les producteurs à lui donner la permission d'archiver leur site. Voir par exemple le résumé sur le site de la British Library : http:// www.bl.uk/aboutus/stratpolprog/legaldep/. Du coup, seules 22 000 captures de 6 000 sites web sont disponibles pour la recherche, un chiffre à comparer aux 2 millions de sites mentionnés par la BNF pour son échantillon représentatif du Web français en 2011, ou même des 30 000 sites concernés par les collectes ciblées, le tout représentant un total de 16,5 milliards de fichiers dans les archives de l'Internet de la BNF fin 2011.

12. Collectes du Web électoral par la BNF, données CSV accessibles en licence ouverte, http:// www.data.gouv.fr/donnees/view/Collectes-du-Web-électoral-par-la-BnF-551866

13. http://gephi.org/

14. C'est le cas des accès par mot-clé ou payants et des liens non HTML, c'est-à-dire notamment les menus déroulants, des liens inclus dans des animations, des liens ouvrant vers des *popups*, de ceux qui n'apparaissent pas explicitement dans la barre d'état du navigateur et des vidéos et sons diffusés en flux plutôt que sous forme d'un fichier à télécharger.

15. Voir par exemple O. Rouchon (CINES) et L. Duplouy (BNF), « *Collaboration over datasets replication* », un résumé disponible en PDF (http://www.mops1.com/oracle/event/pasig/downloads/ Collaboration_over_datasets_replication_CINES_BnF.pdf)

16. http://hal.archives-ouvertes.fr/. HAL est pilotée par le Centre pour la documentation scientifique directe (http://www.ccsd.cnrs.fr). Voir aussi l'atelier *Évolution de l'archive ouverte HAL-SHS* du THATCamp 2012 animé par Laurent Capelli sur HAL_SHS

17. Il fournit aussi à HAL un archivage pérenne de ses collections. http://www.cines.fr/

18. SPAR, projet majeur piloté par la BNF, a été lancé en 2005 et mis en production à partir de 2010 pour permettre à l'institution patrimoniale de rationaliser et mieux gérer son stockage de données numériques. Il aura l'avantage, à terme, de fournir à la BNF un parc de stockage

homogène et des procédures respectant les bonnes pratiques en matière d'archivage de données, tout en offrant une masse critique suffisante pour faire baisser les coûts et mutualiser le stockage entre plusieurs institutions. Voir la présentation de SPAR sur le site de la BNF : http://www.bnf.fr/fr/professionnels/conserver_spar.html

19. http://braceroarchive.org/

RÉSUMÉS

Allons-nous vers une période d'abondance gérée de l'information ou resterons-nous dans la rareté ? L'atelier essaiera d'analyser quelles sont les conséquences de ces pertes de données pour les sciences humaines et sociales, de resituer ces pertes dans un contexte historique (ce n'est pas la première fois que nous perdons des archives) et de dresser un éventail des solutions disponibles, notamment les politiques d'archivage du web actuellement mises en place.

INDEX

Mots-clés : mémoire, archivage, archives, perte de données