RESSI revue electronique suisse de science de l'information

Publiée une fois par année, la Revue électronique suisse de science de l'information (RESSI) a pour but principal le développement scientifique de cette discipline en Suisse.

L'archivage du web : présentation des méthodes de collecte et recommandations pour l'accès aux contenus –et leur structuration-

Ressi — 31 décembre 2016

Jonas Beausire, Haute Ecole de Gestion, Genève

Résumé

Cet article – synthèse d'un travail de bachelor – consiste en l'établissement d'un panorama des grandes approches méthodologiques et stratégies de collecte de l'archivage du web, une analyse des attentes et des résistances du public des chercheurs face à ces nouvelles archives et la présentation de pistes d'innovations et de recommandations pour mieux appréhender l'archivage du web.

Les approches de l'archivage du web sont exposées : intégrale, exhaustive, sélective et thématique. Elles se combinent souvent sur le terrain mais doivent être repensées pour être renouvelées. Chacune d'entre elles peut être accompagnée d'une stratégie de collecte : automatisée, semi-automatisée ou manuelle.

Les attentes des chercheurs, leurs besoins et résistances sont mis en lumière par des résultats d'enquêtes. Si la communauté scientifique s'accorde sur la nécessité de constituer une mémoire du web, la fiabilité et la légitimité des collections issues du web cristallisent les résistances exprimées par les chercheurs. Globalement, les questions épistémologiques et méthodologiques pour inscrire ces archives dans un usage scientifique établi ne sont pas encore résolues.

Enfin, des recommandations techniques et conceptuelles sont abordées : elles mettent notamment l'accent sur la construction d'interfaces d'accès et la description des archives et de leur contexte grâce, en particulier, aux métadonnées. Une variété d'outils d'analyse du web constitue également des leviers privilégiés pour exploiter et mettre en valeur les futures archives du web.

Mots-Clés: Archivage/archive du web patrimoine mémoire chercheur numérique Web archiving/archive heritage memory researcher digital

L'archivage du web : présentation des méthodes de collecte et recommandations pour l'accès aux contenus –et leur structuration-

Introduction

Les questions soulevées par l'archivage du web préoccupent les acteurs du monde de l'information depuis presque vingt ans maintenant. Il est aisé de situer dans le temps les prémices des réflexions qui entourent les questions d'une mémoire du web. En effet, des initiatives comme celle, bien connue et la plus ancienne de toute, de la fondation « Internet Archive »[1] ont pris naissance dès 1996 dans un climat d'urgence à se

saisir des nouvelles traces qui faisaient déjà la mémoire de la fin du XXe siècle (Peyssard 2012). La multiplication des ordinateurs connectés durant la bulle Internet jusqu'en 2000 confirmera la nécessité de sauvegarder les contenus désormais « nés numériques » du prochain millénaire.

L'établissement, durant la première décennie du XXIe siècle, de programmes d'archivage du web institutionnalisés (le plus souvent au sein de bibliothèques nationales) va peu à peu voir opérer un changement de paradigme essentiel : de la numérisation généralisée du patrimoine, il s'est agi de patrimonialiser le (né) numérique. Ce passage, symptôme de la légitimation de ces nouvelles archives, n'est pas sans poser un catalogue de questions : Comment préserver ces nouveaux documents ? Selon quelles logiques documentaires ? Comment les conserver de façon pérenne ? De quelles façons les rendre accessibles et à qui ? Au fond, comment prendre en charge une masse documentaire exponentielle, qui a valeur de patrimoine, et qui ne cesse de disparaître de plus en plus rapidement ?

Si désormais les contenus nés numériques préoccupent les institutions patrimoniales et constituent un segment de notre mémoire collective, les différents acteurs concernés par leur archivage pointent également une autre réalité : la disparition du web d'hier est toujours plus importante. En effet, les documents et données issus du web sont aujourd'hui trop souvent inaccessibles, hantés par le spectre de l'erreur *http 404* (fichier introuvable) ; la durée de vie moyenne d'une page web avant modification ou suppression est d'environ cent jours (Laporte, Kahle, 2011)[2]. La dimension fragile, fuyante et nomade de ces documents exhorte les archivistes, les bibliothécaires et les chercheurs à penser de nouveaux modèles de collecte et plus largement à assimiler de nouveaux lieux de notre mémoire collective.

Les sources mobilisées pour la réalisation de ce travail proviennent majoritairement du web et étonnent souvent par leur complexité lorsqu'elles sont destinées au public de l'ingénierie informatique, par exemple. Le caractère fondamentalement transdisciplinaire des entreprises d'archivage du web se traduit dans la pluralité des publics auxquels s'adressent les ressources scientifiques disponibles sur le sujet. Les publications liées aux activités de l'International Internet Preservation Consortium (IIPC)[3] demeurent le réservoir privilégié des ressources disponibles aujourd'hui, associant souvent études de cas et réflexions holistiques sur l'archivage du web. Les trois axes majeurs de travail du consortium rejoignent ceux associés au circuit du document en bibliothèque : collecte, consultation et préservation (Illien 2011). Par ailleurs, certains médias, spécialisés ou non, soulignent peu à peu les enjeux de la sauvegarde de cette mémoire numérique et cherchent à sensibiliser des publics plus divers et moins avertis.

Au carrefour des enjeux d'accessibilité, de représentativité, de légitimité et de fiabilité des documents nés numériques, cet article[4] se propose de dégager les grandes approches méthodologiques et stratégies de collecte de l'archivage du web à l'œuvre aujourd'hui, mises en regard avec les programmes de la Bibliothèque nationale suisse (BN) et de la Bibliothèque nationale de France (BnF). Il analysera les attentes et les résistances du public des chercheurs face à ces nouvelles archives et enfin présentera des pistes d'innovation et des recommandations pour mieux appréhender l'archivage du web.

Cadre théorique

Il est possible d'identifier aujourd'hui de grandes approches et stratégies de collecte. Une typologie conceptualisée par Thomas Chaimbault (enseignant à l'ENSSIB) offre à voir un panorama des stratégies et modes de dépôt (voir tableau récapitulatif n°1, p. 5) développés par différents établissements nationaux et soutenus par des consortia (Chaimbault, 2008).

En rappelant que ces approches demeurent de purs cadres théoriques et qu'elles doivent être renouvelées – notamment en raison des mutations techniques extrêmement rapides du web – il est également à souligner que les réalités du terrain sont multiples et mêlent bien souvent plusieurs approches et stratégies pour répondre aux besoins spécifiques d'un seul et même programme d'archivage du web.

Les approches

L'approche dite « intégrale » consiste à collecter l'entier du web, sans distinction ni critère de sélection. Les questions liées à une valeur patrimoniale des documents sont évacuées au profit d'un projet de pure exhaustivité. Le projet « Internet Archive » en est aujourd'hui l'exemple unique et concentre la plus large

audience des collections issues du web au travers de son interface d'accès aux documents d'archive, la « Wayback Machine »[5]. Avec près de 273 milliards de pages web collectées (Internet Archive, 2016) et au centre d'un partenariat qui la lie avec près de 440 organismes partenaires, la fondation s'inscrit dans un double mouvement, à la fois celui de la collecte autonome, mais également celui d'un échange continu avec d'autres organisations de collecte. (Leetaru, 2016)

L'approche « exhaustive », quant à elle, vise également une certaine idée de la complétude mais dans un périmètre circonscrit, celui d'un nom de domaine, d'un espace national particulier ou, moins souvent, d'un type de sites. Cette approche, relativement répandue, s'intègre facilement dans les missions d'une institution patrimoniale comme les bibliothèques nationales mais cristallise les ambiguïtés liées à la territorialité du web : des contenus web particulièrement signifiants peuvent être enregistrés sous un nom de domaine hors collecte, par exemple.

A l'inverse des deux précédentes approches, celle dite « sélective » consiste à se saisir de contenus prédéfinis au moyen de critères choisis extrêmement variés : thématiques, en lien avec la nature de la ressource, qualitatifs, etc. Cette approche rompt avec un certain souci d'exhaustivité et se propose de compiler régulièrement des instantanés de sites répondant aux critères de sélection choisis.

Enfin, l'approche « thématique » doit se comprendre comme un embranchement particulier de l'approche sélective : il s'agit d'archiver une collection de sites web en lien avec un événement spécifique. Les collectes des sites web et autres ressources liés aux élections présidentielles françaises menées par la BnF illustrent parfaitement cette approche. (Greffet, 2012) Sa logique renvoie directement à la notion de « collection », voire de « fonds d'archive » puisqu'il s'agit bien pour les bibliothécaires de sélectionner et d'éliminer en vue de former un corpus cohérent, pouvant ainsi former de véritables « produits d'appel » tournés vers un public encore aujourd'hui embryonnaire (Illien, 2008).

Les stratégies

Parallèlement aux différentes approches décrites, Thomas Chaimbault dégage également trois stratégies de collectes différentes. La stratégie « automatisée » qui engage la mise en place d'un logiciel-robot (moissonneur et collecteur du web) : un espace web circonscrit à un domaine choisi est ainsi collecté de façon automatique. Cette stratégie accompagne généralement des approches intégrales ou exhaustives de l'archivage du web. La stratégie « semi-automatisée » implique également l'usage d'un logiciel-robot mais ajoute à son utilisation des critères de sélection plus précis ; elle peut être mobilisée dans le cadre d'une approche sélective du web. Enfin, l'approche « manuelle », même si elle exige également des ressources techniques, replace l'humain au centre des processus de collecte ; les bibliothécaires sont amenés à sélectionner eux-mêmes les sites pertinents. Cette logique combinatoire est essentielle dans le contexte d'une approche thématique.

En assignant la typologie de Thomas Chaimbault au programme d'archivage de la BN, on peut aisément le qualifier de sélectif et thématique. En effet, dans la tradition des Helvetica, la collection « Archives Web Suisse » regroupe en grande majorité des sites web patrimoniaux et helvétiques, selon un périmètre et des critères décidés collégialement. En rejetant tout projet d'exhaustivité et en cartographiant les sites archivés au moyen d'un catalogue de grands principes de sélection et d'exclusion[6], la BN se distingue radicalement de la BnF. En effet, l'institution française combine trois approches : exhaustive, sélective et thématique qui renvoient à autant de modes de collecte. Exhaustive car la BnF procède à des collectes dites « larges » qui moissonnent l'entier du nom de domaine français, mais superficiellement en ce qui concerne la profondeur des sites. Sélective et thématique lorsque la BnF mène ses collectes dites « ciblées » qui visent à s'emparer des sites en profondeur et à une fréquence plus élevée, choisis pour leurs contenus signifiants. (BnF, 2015)

Table 1 : Récapitulatif des grandes approches et stratégies

	Stratégie automatisée	**Stratégie semi-automatisée**	**Stratégie manuelle**
Approche	Entier du web	*Néant*	*Néant*

intégrale	Pas de critère de sélection Logiciel-robot		
Approche exhaustive	Entier du web, mais périmètre précis Nom de domaine ou espace national Logiciel-robot	*Néant*	*Néant*
Approche sélective	*Néant*	Critères de sélection précis Ressources humaines Logiciel-robot	*Néant*
Approche thématique	*Néant*	*Néant*	Critères de la sélection précis Collecte événementielle/thématique Ressources humaines

Un cadre légal différencié, des usages communs

Au cœur du régime différencié des programmes de la BN et de la BnF, réside le cadre légal sur lequel repose les approches mises en œuvre. En effet, ce dispositif structure largement les possibilités des deux institutions. Il est également le reflet d'une « accréditation culturelle de l'éphémère » (Merzeau, 2003). L'absence de dépôt légal suisse au niveau national implique pour la BN une approche sélective et thématique du web. A l'inverse, le dépôt légal du numérique permet à la BnF de s'emparer indifféremment de la quasi-totalité de la production éditoriale numérique française, tendant à une forme d'exhaustivité sans jugement de valeur documentaire. L'accessibilité des archives est une conséquence directe du cadre législatif différent de chacune des deux bibliothèques : la BnF est contrainte d'encadrer son accès pour protéger le droit d'auteur des contenus qu'elle moissonne, alors que la BN est plus souple puisque les accords des producteurs ont été obtenus préalablement.

Il est à noter que l'approche dite « thématique » est partagée par les deux institutions : dans les deux cas, des bibliothécaires sélectionnent en amont les sites signifiants et tentent de former des collections parfois thématiques ou gravitant autour d'événements majeurs. Certains outils informatiques et des préoccupations liées à la profondeur de l'archivage sont partagés par les deux bibliothèques. Le rapprochement s'opère également sur le plan international puisque la BN et la BnF sont membres du Consortium IIPC au travers duquel elles collaborent. Comité de pilotage et groupes de travail au sein de ce consortium sont autant de lieux d'échanges et de retours d'expérience, notamment concernant l'usage de divers logiciels développés par certains membres.

Besoins, attentes et résistances des chercheurs

Le public des nouvelles collections issues des différents programmes de l'archivage du web demeure une question centrale. Si des publics variés peuvent aujourd'hui consulter ces nouvelles archives, celui des chercheurs scientifiques semble être le plus important. (Chevallier, Illien, 2011) (Aubry et al. 2008) De nombreuses questions sont soulevées par ce public particulier : un site Internet peut-il réellement constituer

une source fiable ? Quelle procédure existerait-il pour valider la qualité d'une telle source ? Comment justifier le choix de convoquer tel site plutôt qu'un autre dans une sitographie ?, etc. (Chevallier, Illien, 2011)

Malgré des attentes contradictoires et des réticences notamment méthodologiques et épistémologiques, la communauté scientifique semble s'accorder sur la nécessité de travailler avec le numérique, de s'interroger sur les conditions d'une meilleure appréhension du patrimoine numérique natif et sur la conservation de nouvelles formes d'expressions numériques. (Joutard, 2013) La conservation d'une mémoire du web nécessite une reconceptualisation des modèles traditionnels de l'archivage en les pensant spécifiquement pour les documents numériques natifs. Les pertes de certains contenus nés numériques et les liens morts inquiètent certains acteurs du monde académique, notamment les historiens, qui voient disparaître des ressources à durée de vie limitée. Les doutes et les interrogations se cristallisent majoritairement autour de la fiabilité de ces nouvelles archives dont les contours documentaires peinent à être scientifiquement définis ; en effet, même si certains chercheurs s'accordent autour du bien-fondé de l'archivage des sites institutionnels et des blogs, les actions ou traces individuelles laissées sur le web sont considérées avec davantage de circonspection. (Chevallier, Illien, 2011) C'est bien entendu la question irrésolue de la légitimation du statut de collection de ces archives qui se pose ici en filigrane. Ainsi, l'organisation, la hiérarchisation, voire la discrimination des contenus du web sont attendues de la part des chercheurs pour considérer plus sûrement les nouveaux corpus. La variété des contenus agrégés exige des efforts organisationnels majeurs pour un usage scientifique de ces données. (Leetaru, 2016) La bibliothèque peut endosser un rôle important dans ce processus de légitimation. (Illien, 2011)

Plus concrètement encore, ce sont les difficultés d'accès, autant physiques que techniques, qui préoccupent les chercheurs : le supposé déplacement au sein des bibliothèques dépositaires des fonds et les interfaces difficiles à prendre en main empêchent trop souvent le public de s'approprier ces nouvelles ressources. L'indexation plein texte constitue toujours la voie d'accès privilégiée aux volumes exceptionnels de ces collections, malgré les insatisfactions récurrentes liées aux technologies utilisées (Gomes, Miranda, Costa, 2011). Enfin, l'instabilité du média Internet, la volatilité des données et la difficulté à traiter de gros volumes de données souvent très hétérogènes constituent les principaux freins méthodologiques rencontrés par le monde de la recherche (Mussou 2012).

Recommandations techniques et innovations

En dehors des grandes initiatives nationales et des projets circonscrits à une institution donnée, l'IIPC peut être considéré comme un laboratoire d'innovations incontournable sur la scène internationale. Cet organisme a notamment pour but de sensibiliser aux enjeux liés à la conservation des ressources nées numériques. (Bonnel, Oury, 2014) De nombreuses sources sont accessibles par le biais de cet institut : articles, études de cas et conférences enrichissent des sources souvent disparates sur l'archivage du web. Le lieu des innovations en matière de conservation du web se situe ainsi surtout dans le cadre de collaborations internationales.

Quelques outils du « web vivant »

Il existe aujourd'hui de nombreux outils de pointe pour appréhender, étudier et investir le « web vivant », par opposition au web archivé. Mais comment valoriser, analyser et exploiter des collections issues du web ? Des chercheurs de l'« Oxford Internet Institute »[7] proposent de transposer certains de ces outils aux archives du web (Meyer, Thomas, Schroeder, 2011). Nous en retenons ici quelques-uns :

La visualisation peut constituer une fenêtre d'accès inédite aux archives. Dans l'esprit des infographies, elle permettrait de visualiser la façon dont les différentes archives sont reliées entre elles. Un fort développement de cet outil pour le web vivant existe déjà. La recherche profonde, quant à elle, permet d'interroger finement de gros ensembles de données. La prolifération des informations postées (puis archivées) exigerait ainsi de nouveaux moyens d'accès à de très gros volumes d'informations. Les outils d'analyse des réseaux sociaux (SNA) n'ont pas été adaptés aux archives. Ceux-ci pourraient permettre aux archivistes du web l'analyse des liens hypertextes comme révélateur de la structure des interactions des différents sites web composant leurs collections. Les liens et leur analyse expriment quelque chose de la nature du réseau, de ses jeux d'interactions. Enfin, cette analyse pourrait être complétée par l'archivage des liens et autres annotations qui pointent vers les sites archivés afin d'observer leurs évolutions dans le temps.

D’autres pistes d’innovation sont énoncées au sein de l’étude de Meyer, Thomas et Schroeder (2011) : la première est celle dite du « web cumulatif » : il s’agit de considérer le web archivé littéralement en parallèle du web vivant. Cette organisation en filigrane de couches d’archives viendrait combler la fragmentation et les trous du web (comme les liens morts qui désormais pointeraient vers la ressource archivée). Cette piste, relativement utopique, supposerait un changement structurel et profond du web.

S’il est aujourd’hui possible de comprendre l’organisation et les usages des sites présents sur le web et de consulter certains d’entre eux qui n’existent plus, il demeure impossible encore de comprendre l’usage passé des archives du web. Afin d’y parvenir, les chercheurs proposent d’archiver également les journaux des serveurs (« servers logs ») des sites d’archivage du web. De cette façon, il deviendrait possible de comprendre et d’étudier comment les archives du web ont été ou sont utilisées. En conservant le web d’hier, mais également les usages associés à ce web disparu, il serait possible de combler l’une des attentes exprimées par les chercheurs sur la nécessité d’interroger les pratiques scientifiques qui entourent ces nouvelles archives.

Un autre usage possible des archives du web est celui de ses images et de son fort potentiel visuel. Il s’agirait de saisir certains changements du monde au travers des images circulant sur la toile. En extrayant sur une certaine durée des images d’archives d’un même objet, cela permettrait de superposer les clichés et de proposer un rendu visuel de l’évolution de l’objet.

L’exploitation statistique des archives constitue également une opportunité majeure. Quels sont les outils d’analyse à mettre en place pour faire parler de très importantes collections d’archives du web ? Comment ces outils statistiques permettraient de mieux comprendre la structure des collections et conséquemment celle du web en général ? En s’intéressant, par exemple, aux langues des sites web ou à leur date de création, il serait possible de dégager de grandes tendances structurelles du web. Les travaux d’analyse menés par l’ « Observatoire du dépôt légal » de la BnF[8] sur ses collectes larges s’inscrivent dans cette logique statistique.

En lien avec une analyse structurelle du web, il serait possible de rendre compte de la prolifération d’une idée sur le web, de sa viralité et de ses déplacements. Pour repérer et comprendre où les idées surgissent et comment elles se propagent sur le web, il faut pouvoir remonter à l’origine de l’idée. Cette archéologie suppose une profondeur et une granularité des archives très importantes. Dans cette perspective, la temporalité du web, c’est à dire le tempo des publications et les hyperliens qui les relient, doit être archivée. Sans une profondeur suffisante de l’archivage, cette dimension est impossible à extraire des archives.

Enfin, la question du web illicite interroge les chercheurs sur la meilleure façon de rendre compte de ces matériaux circulant sur le web. Les contenus sexuels illicites, sur les drogues, sur les groupes prônant la haine raciale, le terrorisme, etc. sont nombreux. Quelle entité serait habilitée à prendre en charge leur archivage et dans quel cadre juridique ? Ce genre d’archive pourrait autant intéresser des chercheurs que certaines autorités, la justice ou encore les professionnels de la santé, par exemple. L’enjeu réside ici dans la mise en place d’un mécanisme juridique pour protéger et légitimer l’institution garante de ces documents au contenu illégal, et qui saurait mettre en valeur leur intérêt scientifique. (Meyer, Thomas, Schroeder, 2011)

Interfaces, accessibilité et contextualisation

Malgré de nouvelles perspectives pour l’exploitation de la mémoire du web, les chercheurs constatent une absence globale d’interfaces stables et conviviales pour construire et accéder à de solides archives du web[9]. Dans ces conditions, le déploiement des différentes initiatives demeure compliqué. Plusieurs études (Bonnel, Oury 2014 ; Leetaru, 2012) insistent sur l’opportunité de mettre en place des interfaces d’accès aux archives les plus efficaces possibles, qui sachent explorer de très gros volumes de données. Selon Leetaru (2012), l’interface de Twitter constituerait un modèle standardisé très simple d’utilisation. Ces interfaces doivent également être spécifiquement pensées pour les chercheurs qui formeront sans doute une communauté importante se saisissant de ces archives. Toujours dans la volonté d’offrir une voie d’accès améliorée aux archives, c’est bien une description fine des collections au travers de métadonnées variées qui constituera une mise en valeur des fonds. Cette pratique suppose que les administrateurs des programmes connaissent précisément le contenu de leurs archives, ce qui n’est pas toujours le cas, surtout dans le cadre d’approches exhaustives ou intégrales.

Si les chercheurs se saisissent petit à petit de ces nouveaux contenus et citent désormais des sources provenant de celles-ci, il s'agit donc de penser également à la normalisation de ces citations. Cette préoccupation participe au travail de leur légitimation, qui ne doit pas échapper aux usages actuels de référencement des sources traditionnelles. La mise en place d'un identifiant unique et permanent de chaque page web archivée participerait à un système de citation efficace dans les publications scientifiques. Comme pour la citation des pages du web vivant, certaines métadonnées comme la date de capture de la page sont essentielles pour la constitution de notices complètes. Certaines tentatives basées sur les standards MLA et impulsées par Internet Archive vont dans ce sens aujourd'hui[10].

Si les choix documentaires d'acquisition des bibliothécaires sont longtemps restés opaques pour le grand public, il serait envisageable de renverser cette tendance en documentant les biais, souvent algorithmiques, des crawlers et autres robots qui moissonnent le web pour l'archiver. De la même façon qu'une transparence des politiques documentaires, la mise en lumière de certains détails techniques propres à un programme peuvent contextualiser telle ou telle collection. (Leetaru 2016) Parfois, la date d'archivage d'un site peut ne pas correspondre à la date de capture du même site. Cette réalité peut constituer un biais majeur pour l'étude d'une chronologie exacte de l'évolution d'un site. Si l'on cherche à comparer, par exemple, le nombre de pages traitant de la candidature d'un politicien à une élection avec celui d'un concurrent, les résultats obtenus peuvent ne pas correspondre avec la réalité du web d'alors. Le nombre d'occurrences peut être influencé par certaines politiques d'archivage, par l'algorithme selon lequel le robot moissonne le web, etc. La documentation des archives du web doit éclairer ces potentiels biais techniques. Si on donne à un chercheur la possibilité d'accéder au «journal» du crawler, il pourrait connaître les lieux où le robot a peut-être buté contre tel ou tel contenu : les zones blanches des archives peuvent recéler un sens précieux pour ceux qui les étudient. Par ailleurs, si beaucoup de sites dits « dynamiques » adaptent leurs contenus en fonction de l'emplacement physique de l'internaute, la géographie du robot doit également être un élément de contexte documenté pour les utilisateurs des archives du web ; elle influe directement sur les contenus affichés (et donc collectés), l'ordre des pages, etc. En d'autres termes, un crawler installé en Russie ne collectera pas les mêmes contenus qu'un autre localisé en France.

En définitive, l'ensemble de ces préoccupations techniques renvoie à la question de l'archivage du contexte de l'archive. Les liens sortants d'un site archivés donnent à voir un écosystème global dans lequel le site en question se déploie. Les métadonnées associées ou la localisation du crawler s'inscrivent dans cette même logique. A titre d'exemple, l'archivage des documents audiovisuels du web pratiqué par l'INA suppose l'intégration de paratextes qui vont définir et aider à interpréter et s'approprier les documents d'archive. (Carou, 2007) C'est également une attente spécifique des chercheurs, qui souhaitent pouvoir accéder à toute une série de données contextuelles comme l'URL, la date de capture, la place de la page capturée dans le site, l'arborescence ou encore des statistiques de vue. (Chevallier, Illien, 2011) En conservant le contexte, l'archive fait sens et peut faire rayonner tout son pouvoir mémoriel et remplir sa fonction cardinale de témoignage.

Archives sociales, fonction d'authentification et pédagogie

Suite à l'avènement d'un web ultra collaboratif où les échanges et les commentaires constituent le régime privilégié des internautes, la dimension sociale représente aujourd'hui une part substantielle de l'écosystème global d'un site. Les interactions sociales qui entrent en résonnance avec les documents présents sur le site doivent également être archivées, en mesurant, bien entendu, le perpétuel enrichissement de ses espaces d'interaction. (Meyer, Thomas, Schroeder, 2011)

Les archives du web pourraient également constituer, à terme, un potentiel agent d'authentification. En effet, elles pourraient pointer, par exemple, les changements intervenus sur une page dans un jeu de comparaison entre une page « primaire » (archivée à un moment t) et une page consultée sur le web vivant. Ce travail comparatif prendrait tout son sens dans le contexte mouvant du web. Les pages des sites gouvernementaux ou médicaux et leurs évolutions pourraient ainsi être authentifiées par les archives qui, une fois de plus, rempliraient leur objectif de garantie d'authenticité du document. (Meyer, Thomas, Schroeder, 2011)

Enfin, afin de sensibiliser les plus jeunes aux enjeux de l'archivage du web, des programmes impliquant des élèves dans l'élaboration de collections d'archive du web ont été mis en place.[11] (Reynolds, 2013) Il s'agit de rendre attentives les futures générations à l'importance de ce patrimoine nouveau. Les « digital natives »

doivent prendre conscience que les contenus produits sur le web ne sont pas éternels et qu'une importante partie de notre mémoire collective se crée, circule et meurt parfois sur la toile. Gageons que cette génération, si active sur le web et coutumière de la richesse de ses contenus, mesurera plus facilement les enjeux d'une sauvegarde de la mémoire numérique que ses aînés.

Conclusion

A ce jour, quatre grandes approches de l'archivage du web peuvent être identifiées : intégrale, exhaustive, sélective et thématique. Chacune d'entre-elles peut parfois être accompagnée d'une stratégie de collecte particulière : automatisée, semi-automatisée ou manuelle. Ces différentes approches constituent des cadres théoriques qui se combinent parfois sur le terrain et qui doivent se renouveler et s'adapter notamment à de nouvelles réalités de l'archivage des documents nés numériques.

Les chercheurs ont tout à la fois des attentes et des résistances : quoiqu'issus d'horizons disciplinaires différents, ils s'accordent sur la nécessité de conserver une mémoire du web, alors même que la disparition des documents nés numériques inquiète certains d'entre eux. C'est autour d'une politique documentaire pensée pour former des collections qui n'apparaissent pas toujours comme légitimes ou fiables aux yeux des chercheurs que la sélection des contenus doit s'articuler. La difficile prise en main des interfaces d'accès à ces archives doit être résolue pour faire rayonner toute la richesse de ses contenus. Un échange des usages et des compétences à l'international peut y répondre, comme on le constate au sein de l'IIPC.

Des outils d'analyse du web vivant comme la visualisation des contenus, la recherche au sein de gros ensembles de données ou l'analyse des réseaux sociaux, sont autant de leviers à activer et transposer pour exploiter et mettre en valeur les collections des archives du web. D'autres pistes d'innovations, comme l'archivage des journaux des serveurs pour comprendre l'usage passé des archives, l'exploitation statistique des archives ou encore l'observation de la prolifération d'une idée sur le web forment un avenir prometteur des archives du web. Les interfaces d'accès constituent à la fois les vitrines des collections et les portes d'accès principales aux contenus ; exploratrices de gros volumes de données, elles doivent être sans cesse repensées pour garantir un accès toujours plus facilité. Le travail de description des archives et des robots-crawler et l'inscription systématique de métadonnées sont des recommandations centrales et récurrentes dans les études. L'archivage du contexte de ces archives répond aux attentes des chercheurs et tend à inscrire ces nouveaux corpus dans une tradition théorique archivistique, notamment concernant leur fiabilité.

L'établissement d'une mémoire numérique apparaît sinon urgent, du moins légitime. Il demeure plus que jamais nécessaire de poursuivre les efforts de recherche autour des nombreuses questions posées par les programmes d'archivage du web : la complexité des processus à mettre en œuvre, les innovations technologiques associées, les politiques et les choix documentaires, mais également la place des professionnels de l'information dans les mécaniques d'archivage sont des enjeux majeurs.

En concentrant un maximum les actions du quotidien d'une société sur son réseau, Internet tend à devenir un lieu de notre histoire mondiale. La trace, le signe ou l'indice numérique nous invitent plus que jamais à considérer le web et son archivage comme une véritable archéologie des pratiques humaines.

Notes

[1] Pour davantage d'informations sur le projet et pour notamment accéder à la « Wayback machine », consulter : https://archive.org/index.php

[2] D'autres chercheurs présentent des chiffres moins alarmistes mais néanmoins préoccupants : 80% des pages sont mises à jour ou disparaissent après un an. (Gomes, Miranda, Costa, 2011)

[3] Pour davantage d'informations sur cet organisme international, consulter : http://netpreserve.org/about-us

[4] Il constitue une synthèse du travail de bachelor intitulé « L'archivage du web : stratégies, études de cas et recommandations », disponible dans son intégralité à l'adresse : https://doc.rero.ch/record/257793?ln=fr

[5] A noter que cette interface d'accès permet d'accéder uniquement à un nombre restreint de ressources. En effet, une grande partie des contenus reçus ou collectés par la fondation ne sont que partiellement accessibles en raison d'embargos, contrats de licence et autres politiques d'accès. (Leetaru, 2016)

[6] L'ensemble de ces grands principes de sélection et d'exclusion est disponible au sein du document consultable ici : https://www.nb.admin.ch/nb_professionnel/01693/01699/01873/01895/index.h...

[7] Pour davantage d'informations sur cet institut, consulter : http://www.oii.ox.ac.uk/

[8] Pour davantage d'information sur cet observatoire, notamment les rapports produits, consulter : http://www.bnf.fr/fr/professionnels/depot_legal_definition/s.depot_legal_observatoire.html?first_Art=non

[9] Il est à noter néanmoins que les interfaces ne cessent de s'améliorer pour mieux s'adapter à leurs utilisateurs, comme en témoigne la récente mise à jour de celle de la BnF en octobre 2016.

[10] Consulter à ce propos : http://www.writediteach.com/images/Citing%20from%20a%20Digital%20Archive%20like%20the%20Internet%20Arc

[11] C'est le cas, par exemple, de l'initiative « K-12 Web Archiving » : https://archive-it.org/k12/

Bibliographie

AUBRY, Sara et al., 2008. Méthodes techniques et outils. Documentaliste-Sciences de l'Information [en ligne]. Avril 2008. Vol. 45. p.12-20. [Consulté le 01.11.2016]. Disponible à l'adresse : http://www.cairn.info/revue-documentaliste-sciences-de-linformation-2008-4-p-12.htm

BNF, 2015. Collectes ciblées de l'internet français. Bnf.fr [en ligne]. 26 mars 2015. [Consulté le 01.11.2016]. Disponible à l'adresse : http://www.bnf.fr/fr/collections_et_services/anx_pres/a.collectes_ciblee.... Html

BONNEL, Sylvie, OURY, Clément, 2014. La sélection de sites web dans une bibliothèque nationale encyclopédique : une politique documentaire partagée pour le dépôt légal de l'internet à la BnF. IFLA World Library and Information Congress 80th IFLA General Conference and Assembly, Lyon, 16-22 August 2014 [en ligne]. [Consulté le 01.11.2016]. Disponible à l'adresse : http://library.ifla.org/998/1/107-bonnel-fr.pdf

CAROU, Alain, 2007. Archiver la vidéo sur le web : des documents ? Quels documents ?. Bulletin des bibliothèques de France [en ligne]. 2007. N°2. [Consulté le 01.11.2016]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2007-02-0056- 012

CHAIMBAULT, Thomas, 2008. L'archivage du web [en ligne]. Dossier documentaire. Villeurbanne : enssib. 2008. [Consulté le 01.11.2016]. Disponible à l'adresse : http://www.enssib.fr/bibliotheque-numerique/documents/1730-l-archivage-du-web.pdf

CHEVALLIER, Philippe, ILLIEN, Gildas, 2011. Les Archives de l'Internet : une étude prospective sur les représentations et les attentes des utilisateurs potentiels [en ligne]. Bibliothèque nationale de France. 2011. [Consulté le 01.11.2016]. Disponible à l'adresse : http://www.bnf.fr/documents/enquete_archives_web.pdf

GOMES, Daniel, MIRANDA, Joao, COSTA, Miguel, 2011. A survey on web archiving initiatives. In: *International Conference on Theory and Practice of Digital Libraries* [livre électronique]. Berlin, Springer, pp. 408-420. Lecture Notes in Computer Science, 6966. [Consulté le 01.11.2016]. Disponible à l'adresse: http://sobre.arquivo.pt/about-the-archive/publications-1/documents/a-survey-on-web-archiving-initiatives

GREFFET, Fabienne, 2012. Le web dans la recherche en science politique [en ligne]. Revue de la Bibliothèque nationale de France [en ligne], n°40. 2012. [Consulté le 01.11.2016]. Disponible à l'adresse : www.cairn.info/load_pdf.php?ID_ARTICLE=RBNF_040_0078

ILLIEN, Gildas, 2011. Une histoire politique de l'archivage du web. Bulletin des bibliothèques de France [en ligne], n°2, 2011. [Consulté le 01.11.2016]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012

ILLIEN, Gildas, 2008. Le dépôt légal de l'internet en pratique. Bulletin des bibliothèques de France [en ligne], n° 6, 2008. [Consulté le 01.11.2016]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2008-06-0020-004

INTERNET ARCHIVE, 2016. *Archive.org* [en ligne]. [Consulté le 01.11.2016]. Disponible à l'adresse : https://archive.org/web/

JOUTARD, Philippe, 2013. Révolution numérique et rapport au passé. Le Débat [en ligne], n°177, 2013. [Consulté le 01.11.2016]. Disponible à l'adresse : http://www.cairn.info/revue-le-debat-2013-5-page-145.htm

LAPORTE, Xavier, KAHLE, Brewster, 2011. Brewster Kahle, Internet Archive: "Le meilleur du web est déjà perdu". *Internet Actu* [en ligne]. Mars 2011. [Consulté le 01.11.2016]. Disponible à l'adresse: http://www.internetactu.net/2011/06/28/brewster-kahle-internet-archive-le-meilleur-du-web-est-deja-perdu/

LEETARU, Kalev H., 2012. A vision of the role and future of web archives. IIPC 2012 General Assembly, [en ligne], 2012. [Consulté le 01.11.2016]. Disponible à l'adresse : http://netpreserve.org/sites/default/files/resources/VisionRoles.pdf

LEETARU, Kalev H., 2016. The Internet Archive Turns 20 : A Behind The Scences Look At Archiving The Web. *Forbes* [en ligne]. 18 janvier 2016. [Consulté le 01.11.2016]. Disponible à l'adresse : http://www.forbes.com/sites/kalevleetaru/2016/01/18/the-internet-archive-turns-20-a-behind-the-scenes-look-at-archiving-the-web/#172e34fb7800

MERZEAU, Louise, 2003. Web en stock. *Cahier de médiologie* [en ligne]. 2003. P. 158-167. [Consulté le 01.11.2016]. Disponible à l'adresse : https://halshs.archives-ouvertes.fr/halshs-00487319

MEYER, Eric T., THOMAS, Arthur, SCHROEDER, Ralph, 2011. Web Archives : The Future(s). IIPC netpreserve.org [en ligne]. 2011. [Consulté le 01.11.2016]. Disponible à l'adresse : http://netpreserve.org/sites/default/files/resources/2011_06_IIPC_WebArchivesTheFutures.pdf

MUSSOU, Claude. Et le web devint archive : enjeux et défis. Ina-expert.com [en ligne]. Juin 2012. [Consulté le 01.11.2016]. Disponible à l'adresse : http://www.inaexpert.com/e-dossier-de-l-audiovisuel-sciences-humaines-et-sociales-et-patrimoinenumerique/et-le-web-devint-archive-enjeux-et-defis.html

PEYSSARD, Jean-Christophe, GINOUVES, Véronique, 2012. Internet Archive. Aldebaran.revues.org [en ligne]. 2 septembre 2012. [Consulté le 01.11.2016]. Disponible à l'adresse : http://aldebaran.revues.org/6339

REYNOLDS, Emily, 2013. Web Archiving Use Cases. Library of Congress, UMSI, ASB13 [en ligne]. Mars 2013. [Consulté le 01.11.2016]. Disponible à l'adresse : http://netpreserve.org/sites/default/files/resources/UseCases_Final_1.pdf

‹ Editorial haut Bibliothèques et quête d'identité ›

Version imprimable Vous devez vous connecter pour poster des commentaires Publié par Ressi