L'archivage des éphémères sur le web : les paradoxes d'une mission patrimoniale

Sophie Derrot

Colloque « Les éphémères, un patrimoine en devenir »
Janvier 2014

Patrimèph

L'archivage des éphémères sur le web : les paradoxes d'une mission patrimoniale

Sophie Derrot

La question de l'éphémère est prégnante dans la considération d'un média aussi souple, aussi réactif, aussi vivant que le web. Chacun d'entre nous a fait l'expérience d'un marque-page soigneusement enregistré dans son navigateur qui, d'une consultation à l'autre, n'aboutit plus sur rien, d'une page d'erreur 404 lors d'une navigation ou bien d'un site qui a changé d'adresse (avec ou sans redirection). L'éphémère fait partie de la définition du web tel que nous le connaissons aujourd'hui. Les contenus en ligne semblent même s'orienter dans une part croissante vers un caractère éphémère volontaire.

Le web est un média de flux. Il peut alors paraître paradoxal, voire contradictoire de tenter de figer ces contenus dans un état historique et de les patrimonialiser tels quels : pourquoi conserver ce moment plutôt qu'un autre ? Mais le web est aujourd'hui aussi l'un des vecteurs de notre culture, qu'elle soit scientifique ou quotidienne, institutionnelle ou populaire. Il est donc évident que la question de la conservation d'un tel témoignage se pose dès aujourd'hui, particulièrement dans un cadre comme celui du dépôt légal.

Historique de l'archivage du web

L'internet est un média jeune en comparaison d'autres qui ont été abordés lors de ces journées sur le patrimoine éphémère. Il a en effet été rendu progressivement accessible au grand public à partir du début des années 1990. Cependant, l'idée de garder une trace de ces premières publications en ligne est apparue assez rapidement : on a toujours la capture d'écran de la première page publiée en ligne, en 1991[1]. Pour les opérations massives d'archivage de sites web, il faut toutefois attendre le milieu des années 1990, avec la création en 1996 de la fondation californienne à but non lucratif Internet Archive, ainsi que les initiatives australienne (PANDORA) et canadienne (EPPP).

En France, la Bibliothèque nationale et l'Institut national de l'audiovisuel mènent plusieurs études dès la fin des années 1990. Les premières opérations volontaires de capture menées par la BnF ont lieu en 2002, autour des élections présidentielles, avec la collaboration d'Internet Archive et bien avant la mise en place d'un cadre légal et organisationnel. La mission de dépôt légal est étendue à l'internet en 2006 par la loi sur les droits d'auteur et droits voisins dans la société de l'information (DADVSI) et son décret d'application, publié en 2011. Les institutions dépositaires de ces nouveaux contenus sont la BnF et l'INA, suivant leurs périmètres d'action traditionnels : les contenus concernant les chaînes de radio et télévision pour l'INA, le reste pour la BnF. Ce dépôt légal a aujourd'hui sa place dans le Code du patrimoine, comme le reste des textes liés à cette mission multiséculaire. Il se place en effet dans la suite des adaptations successives du dépôt légal institué par François Ier (1537) aux différents supports de la production intellectuelle française, selon une évolution logique puisqu'une partie de cette production n'est disponible aujourd'hui qu'en ligne.

1. Le CERN a republié cette page à l'adresse : http://info.cern.ch/hypertext/WWW/TheProject.html (consultée le 1er avril 2015)

Cependant, des adaptations du cadre traditionnel ont été nécessaires pour prendre en compte les particularités de ces « signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique » cités par la loi. Une de ces mises à jour notables est l'inversion du rapport entre déposant et institution dépositaire : ce n'est plus au producteur des contenus publiés de les déposer, mais à l'institution de les collecter en ligne. Si le seuil minimal fixé par la loi est d'une collecte de ces contenus par an, les deux institutions concernées évaluent les fréquences nécessaires selon leurs propres critères et selon les rythmes de publication. La loi précise également le périmètre d'action de ce dépôt légal, celui d'un « internet français[2] ». Enfin, un ajustement d'importance a transformé la position de principe de ce nouveau dépôt légal, non plus basée sur l'idéal traditionnel d'exhaustivité, mais sur la représentativité et de l'échantillonnage raisonné. Une reconnaissance est faite par-là du caractère massif et éphémère de ces nouveaux contenus et de la difficulté de s'y confronter dans une démarche patrimoniale.

Constituer une collection représentative d'une telle masse et de temporalités aussi variables est un défi nécessaire. Mais le paradoxe intervient alors : comment rester dans une absence de jugement de valeur tout en veillant à l'équilibre des collections ?

Une organisation spécifique

Un nouveau modèle d'organisation s'est mis en place au sein de la BnF pour collecter ces contenus et les intégrer aux collections de la bibliothèque. L'équipe chargée de cette mission est constituée à fois de personnels des métiers des bibliothèques et d'informaticiens, développeurs et ingénieurs de production. Ils sont secondés par plusieurs dizaines de robots de collecte, logiciels indispensables puisque la BnF récupère elle-même la plupart des contenus directement en ligne. Le robot (Heritrix 1.14) se déclare comme un navigateur classique et se comporte comme un internaute automatique. Il capture tous les contenus proposés au visiteur du site concerné à l'instant précis de sa visite. Le contenu est alors enregistré sous cette forme et son code est figé. Il entre ainsi dans les collections, marqué d'un tampon virtuel de dépôt légal précisant la date exacte de la collecte. L'inscription de la dimension chronologique dans les archives du web ainsi collectées est indispensable afin de donner un sens historique à ces contenus, pour reconstituer ensuite dans l'interface d'accès l'évolution, la vie (voire la mort) de la publication en question.

Pilier de la mission de dépôt légal de la BnF, la préservation de ces collections, qui adoptent les formats les plus divers, doit être pensée dès la collecte. Les archives du web reflètent le caractère hétéroclite des formats de publication en ligne (sites web en HTML ou en flash, fichiers PDF, DOC, fichiers image, vidéo ou audio de formats divers, etc.), ce qui en fait la richesse et la représentativité. Cependant, cette diversité porte en elle le risque d'obsolescence de certains de ces formats sur le long terme, accru par la rotation rapide des contenus et de leurs formes de diffusion. Le système de préservation qui a été mis en place à la BnF permet de se confronter à une grande masse de données – les archives du web représentaient 567 To de données au 1er janvier 2015 et s'augmentent actuellement de plus de 100 To par an – et la nécessité absolue de conserver et de documenter le contexte de collecte et ce que l'on a pu capturer du cadre de publication, comme les liens entre les pages, les formats, etc.

La collecte et la conservation au titre du dépôt légal entraînent de facto une patrimonialisation des contenus, mais ne suffisent pas à la complétude de la mission. La consultation doit être offerte aux publics intéressés, notamment scientifiques. En contrepartie à la grande liberté ménagée par le cadre juridique, l'accès aux contenus archivés est restreint aux espaces Recherche de la BnF et de ses partenaires en région[3]. Ces espaces, accessibles

2. Selon le décret d'application de 2011, sont définies comme appartenant au champ d'obligation d'un dépôt légal les publications sur des sites enregistrés en France, les publications éditées par des personnes physiques ou morales domiciliées en France, ou les publications produites sur le territoire national, même si elles sont diffusées par une société étrangère.

3. Ces partenaires sont les bibliothèques de dépôt légal imprimeur, têtes de pont en région. Au mois d'avril 2015, l'accès aux archives de l'internet est possible à la médiathèque Émile Zola de Montpellier,

uniquement sur accréditation, proposent une interface spécifique pour la consultation, coupée du réseau : il ne s'agit pas de l'internet mais bien d'une archive, figée, avec une chronologie interne, qui reconstitue l'environnement de lecture et de navigation original au moment du passage du robot sur le site capturé. La date de l'archive consultée peut être choisie sur un calendrier proposant toutes les captures du site désiré, mettant en avant la dimension temporelle de l'archive. Une fois sur le site à la date choisie, il est possible pour le chercheur de citer précisément la page consultée grâce à un permalien[4] précisant la source du document (les archives de l'internet de la BnF), sa date précise de capture et son emplacement d'origine (URL de la page capturée). Ce permalien permet la citation exacte de manière pérenne, y compris la temporalité d'une ressource dont la présence a pu être fugitive en ligne.

Les grands types de collectes

Les collectes ont lieu selon des contextes documentaires différents, adaptés au mieux aux rythmes de publication des contenus. Trois grands axes président à l'organisation de ces captures.

Le premier type de collecte est la collecte annuelle du domaine français, appelée « collecte large », qui couvre l'obligation de dépôt légal au sens strict : il s'agit de capturer une part représentative des sites web définis comme appartenant au périmètre de la BnF, sur une période assez courte pour que la collection conserve une cohérence temporelle. La capture de cet instantané du web français se déroule une fois par an depuis 2004 et elle est menée en interne à la BnF depuis 2010. La capture des sites faite à cette occasion est peu profonde (quelques milliers d'URL sont collectées par domaine), mais elle englobe la plus grande quantité de sites possible : chacun des 4,2 millions de sites de la collecte large de 2014 a ainsi été collecté selon les mêmes critères, pour un résultat de 67,32 To de données. Cette collecte poursuit un objectif de représentativité globale de l'internet français, selon les principes du dépôt légal, sans principe de sélection documentaire liée aux contenus des sites. Cette collecte large permet d'ailleurs de capturer des sites qui sinon n'entreraient pas dans les collections, car n'appartenant pas au périmètre de la politique documentaire de la BnF, comme les sites commerciaux de vente à distance ou les forums.

Le deuxième type de collecte, les collectes dites « ciblées », s'appuie sur le travail d'une centaine de correspondants des départements de collection et de bibliothèques partenaires de la BnF en région ; ceux-ci sélectionnent des sites en lien avec leur domaine de spécialité, dans une optique de complément et de continuité des collections. Certains contenus entrent de moins en moins dans les collections sous leur forme papier (publications officielles, programmes de théâtre[5]), d'autres apportent une valeur ajoutée indéniable sous leur forme en ligne. 40 000 sites environ sont collectés dans ce cadre. Ces collectes ciblées complètent la précédente, car leurs paramètres techniques sont plus souples et permettent notamment une adaptation fine de la fréquence et de la profondeur de capture : certains contenus sont particulièrement éphémères ou demandent une attention particulière (technique, documentaire). Il en est ainsi des sites de manifestations, comme celui du festival de bandes dessinées d'Angoulême, dont la fréquence de collecte est réglée sur celle du festival.

Le dernier axe de collecte suit une logique plus événementielle ou thématique et concerne des projets spécifiques. Elles sont souvent plus ponctuelles, avec une politique documentaire définie, orchestrée par plusieurs départements de la bibliothèque, souvent en coopération avec des acteurs extérieurs.

à la bibliothèque nationale et universitaire de Strasbourg et à la bibliothèque Stanislas de Nancy. Cet accès a vocation à s'étendre dans les années qui viennent.

4. De type http://archivesinternet.bnf.fr/20081118095012/http://expositions.bnf.fr/japonaises/index.htm

5. Voir l'article de Iris Berbain et Cécile Obligi, « Conserver l'éphémère du théâtre, du programme de spectacle au site web », dans *hybrid, revue des arts et médiation humaines*, 2014/1, disponible à l'adresse : http://www.hybrid.univ-paris8.fr/lodel/index.php?id=187 (consulté le 8 avril 2015).

Elles peuvent viser un type de ressources particulier, comme la presse : depuis 2010, les pages d'accueil et les articles de une d'une centaine de sites web d'actualité en ligne sont collectés tous les matins, qu'il s'agisse de presse nationale (lemonde.fr, liberation.fr) ou régionale (ouest-france.fr, leprogres.fr), ayant un équivalent papier (20minutes.fr, present.fr) ou non (mediapart.fr, rue89lyon.fr). Cette collecte de l'actualité au jour le jour a déjà montré sa pertinence par les multiples changements qu'a connus la presse ces dernières années : elle témoigne ainsi de l'apparition en ligne de *France soir* en 2004 et sa disparition en 2012, ou bien des fréquents changements esthétiques et organisationnels de la page d'accueil du site lemonde.fr depuis 1997.

Ces collectes « projets » peuvent également concerner des thèmes spécifiques (science-fiction, mouvements sociaux) ou bien des événements. Dans cette dernière acception, les collectes projets trouvent leur raison d'être dans le caractère éphémère des contenus qu'elles visent à capturer : jeux olympiques, révolutions arabes, élections. Les contenus électoraux font l'objet d'une attention particulière à chaque grand scrutin depuis 2002[6]. En effet, les sites politiques sont pleinement concernés par l'évolution rapide des contenus et leur disparition, même dans le cas de sites institutionnels, comme celui de l'Élysée. Depuis 13 ans, le corpus ainsi constitué documente de façon inédite l'évolution de la pratique politique en ligne et la cristallisation du débat dans ce nouvel espace[7], avec un cadre de sélection suivant une typologie stable depuis 2007, qui couvre les sites officiels de candidats ou de partis, mais également des sites d'observatoires ou de la société civile. Les sites de partis par exemple témoignent d'une migration en ligne de contenus éphémères traditionnellement imprimés (programme, flyers, affiches), souvent dans un but de plus grande dissémination (comme le Parti socialiste et sa « valise du militant » à imprimer en 2012). Ces collectes sont également l'occasion de capturer ce qui fait l'éphémère politique sur le web, comme l'activité sur Twitter, particulièrement importante en 2012 — 1 090 comptes ont alors été capturés à intervalle régulier.

L'organisation et le suivi de ces collectes projets, particulièrement des collectes électorales, a permis à l'équipe du dépôt légal de l'internet de la BnF d'acquérir une expérience et une souplesse précieuses. Ainsi, la BnF est à même aujourd'hui de capturer rapidement et de manière unitaire des sites menacés de disparitions ou qui font ponctuellement l'actualité. Cette organisation en « collecte d'urgence » permet tout autant de collecter les réactions face à des événements soudains (autour des attentats des 7 et 8 janvier 2015) que des épiphénomènes particulièrement liés au médium internet (les Tumblr de gifs animés qui ont fleuri début 2012).

La question de l'éphémère sous-tend l'activité de dépôt légal du web et de collecte du patrimoine intellectuel qui circule sur l'internet. L'injonction de représentativité, confrontée à la problématique de la masse des contenus concernés, trouve actuellement des solutions dans l'intervention paradoxale d'une politique documentaire. Celle-ci permet justement de s'adapter aux rythmes de publication et aux types de contenus qui ont vocation à rejoindre les archives de l'internet de la BnF. Il est également central de documenter les pratiques professionnelles qui ont permis la collecte de ces documents ; principe d'activité du robot, choix techniques de préservation, clarté de l'interface de consultation : la possibilité d'expliquer précisément le contexte de collecte est indispensable à une bonne compréhension de ces collections. Il reste cependant que la volonté poussée d'éphémère qui parcourt aujourd'hui certaines pratiques du web (les applications comme Snapchat) peut faire échec à cette mission de dépôt légal, qui se doit donc d'adopter souplesse et pragmatisme.

6. Présidentiel et législatif en 2002, 2007 et 2012 ; régional en 2004, 2010 et 2015 ; européen en 2004, 2009 et 2014.

7. Sur la collecte des contenus politiques, voir Clément Oury, « Soixante millions de fichiers pour un scrutin. Les collections de sites politiques à la BnF », dans *Revue de la BnF*, 2012/1, n° 40.