Le dépôt légal de l'internet à la Bibliothèque nationale de France

Clément Oury

Le dépôt légal de l'internet à la Bibliothèque nationale de France : entre représentativité et sélection documentaire

Clément Oury (BnF)

La maturation d'un « web patrimonial »

Depuis le milieu des années 1990, c'est-à-dire quelques années seulement après la naissance du web, se développent les principes et les méthodes d'un archivage à des fins patrimoniales[1] : les publications en ligne, au statut éphémère et fragile, sont autant de ressources qu'il s'agit de faire entrer dans des collections publiques et de préserver à destination des générations futures… ou de ceux qui voudront, d'ici quelques années, consulter les documents qui auront déjà disparu.

À la Bibliothèque nationale de France (BnF), cet objectif est poursuivi dans le cadre pluriséculaire du dépôt légal qui, dès 1537, édicte que toute publication produite ou diffusée en France doit entrer dans les collections nationales. Depuis cinq siècles, ce dispositif juridique s'est adapté aux différentes évolutions du monde éditorial : après les imprimés, les estampes, le son, la vidéo, ou encore les logiciels se sont vus soumis au dépôt. Les premières réflexions relatives à la prise en compte du caractère patrimonial de l'internet en France datent de la fin des années 1990, sous l'influence d'organismes novateurs comme la fondation américaine Internet Archive ou les bibliothèques nationales de Suède et d'Australie. En 2002, la BnF collecte les sites relatifs aux élections qui voient la victoire de Jacques Chirac et l'éviction de Lionel Jospin. En parallèle de ces expérimentations techniques, la loi mûrit lentement : enfin, le 1er août 2006, le Parlement se prononce en faveur d'un « dépôt légal des signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique »[2].

Au vu de l'immensité du périmètre à couvrir, deux institutions sont chargées de ce dépôt, dans la continuité de leurs missions respectives : l'Institut national de l'audiovisuel (INA) se voit confier les sites de radio et de télévision, à la BnF revient l'ensemble des autres sites de l'internet français. Ce dispositif juridique, désormais intégré dans le code du patrimoine, permet à ces institutions patrimoniales d'archiver l'ensemble des publications en ligne. En revanche, il comporte un important revers du point de vue de l'accès : pour des raisons de respect de la propriété intellectuelle, mais aussi par souci de protection des données personnelles, les collections sont uniquement accessibles dans les salles de recherche des établissements dépositaires ainsi que de leurs principaux partenaires en région.

Le dépôt légal de l'internet : un cycle complet, de la sélection à la conservation

Ainsi, en matière juridique, c'est la continuité qui prime. Cependant, les modes d'entrée changent : même si l'on parle toujours de dépôt, les organismes dépositaires mettent en œuvre des procédures de collecte. La BnF comme l'INA utilisent la technologie des « robots » : il s'agit en fait de logiciels qui, à partir d'une liste d'adresses URL qui leur est indiquée, parcourent le web de lien en lien pour découvrir et capturer les contenus qu'on les a chargés de moissonner.

À la BnF, les données collectées sont ensuite indexées pour permettre une navigation « temporelle » dans les archives du web. En saisissant l'adresse URL du site que l'on recherche, le lecteur peut accéder à ses différentes strates, c'est-à-dire à son état aux différents moments où il a été capturé. Pour une date choisie, il peut ensuite effectuer une navigation « spatiale », en surfant sur les sites voisins comme l'aurait fait un internaute à l'époque – à condition bien entendu que les contenus demandés aient été capturés.

---

[1] Les publications relatives au dépôt légal de l'internet à la BnF sont recensées dans une « bibliographie sélective » : http://www.bnf.fr/documents/bibliographie_dl_web.pdf.

[2] Voir www.bnf.fr/fr/professionnels/depot_legal_definition/i.depot_legal_loi/s.depot_legal_loi_code.html.

Enfin, les ressources archivées sont versées dans l'entrepôt de préservation numérique de la BnF, le système SPAR (système de préservation et d'archivage réparti). Pour garantir la sécurité « physique » des collections numériques, les données sont copiées à l'identique sur plusieurs sites distants ; d'autre part, le format dans lequel elles sont encodées est identifié automatiquement et cette information est conservée pour permettre les futures opérations de préservation – précaution essentielle tant les formats des fichiers sur le web sont multiples et susceptibles d'obsolescence.
La plupart des outils utilisés par la BnF ont été développés dans le cadre du consortium international pour la préservation de l'internet : l'IIPC, fondé en 2003, regroupe aujourd'hui une cinquantaine d'institutions patrimoniales et de recherche sur les cinq continents. Le développement en coopération des outils nécessaires à l'archivage du web est effectivement l'un des objectifs majeurs du consortium. L'échange de principes et de bonnes pratiques en matière de sélection documentaire en est un autre ; dans ce domaine en revanche, l'uniformité n'est pas de mise car chaque institution choisit son propre modèle d'archivage.

Les paradoxes d'un dépôt légal partiellement sélectif

En matière de web, l'exhaustivité – but originel du dépôt légal – n'est plus un objectif accessible : il n'est pas possible de capturer chaque site à chaque mise à jour. La BnF vise donc la représentativité : il s'agit de constituer une image, un « instantané » de l'internet français, qui prenne en compte tous les types de publications (du site officiel à la plate-forme de diffusion de vidéos ou aux parties publiques des réseaux sociaux), et tous les contenus, du plus sérieux au plus dérisoire.
À cette fin, la BnF conjugue deux modèles de collecte : le premier est la collecte « large », réalisée une fois par an, qui concerne tous les sites qui ont été automatiquement identifiés comme français – soit plus de quatre millions à ce jour. Les collectes « ciblées » en revanche concernent des sites à capturer plus fréquemment (jusqu'à une fois par jour) ou plus profondément (jusqu'à plusieurs centaines de fichiers par domaine) ; il peut également s'agir de ressources à collecter en raison de leur lien à un événement donné (élections, festivals, rencontres sportives…). Ces sites-là, près de trente mille à ce jour, sont identifiés individuellement, soit par des bibliothécaires de la BnF (une centaine d'agents de la BnF, répartis dans les différents départements thématiques, participent à la sélection), soit par des partenaires (bibliothèques, centres d'archives, laboratoires de recherche). Prenons l'exemple de la littérature contemporaine : le département Littérature et art a identifié des sites de référence en matière d'analyse ou de critique, il a également collaboré avec l'Association pour l'autobiographie[3] pour recenser plusieurs centaines de blogs, équivalents en ligne des journaux personnels[4].

Le « modèle intégré » adopté à la BnF vise donc à conjuguer les avantages de la logique du dépôt légal (constituer un « miroir » de la production et de la consommation culturelles françaises) et ceux de la sélection documentaire (conserver les pans les plus dynamiques et novateurs de l'internet). Comparée à ses homologues internationaux, la BnF se situe donc à mi-chemin, certaines institutions reposant exclusivement sur des collectes automatiques, tandis que d'autres ne conservent que les segments qu'ils jugent les plus essentiels de leur web national. L'expérience de l'archivage du web montre ainsi qu'au-delà des moyens techniques et des ressources mis à disposition, ce sont le cadre juridique, les missions et les traditions scientifiques qui priment dans la constitution des collections nationales.

[3] http://autobiographie.sitapa.org/.
[4] Sur ce sujet, voir Gildas Illien, « Les mémoires de la Toile, l'archivage d'Internet à la BnF », dans *La Faute à Rousseau*, n° 45, 2007.