

L'AUTEUR



Gildas ILLIEN est conservateur et chef du Service Dépôt légal numérique à la Bibliothèque nationale de France, à Paris.

l'expérience du furetage dans un grenier familial, voire... dans une bibliothèque.

Alors, garder la trace du Web aurait peut-être et malgré tout quelque chose de bon ? Une histoire tristement ordinaire et contemporaine l'illustre. Symbole de l'industrie dans l'Est de la France, l'entreprise *Vogica*, spécialiste depuis 1976 des cuisines et salles de bain faites sur mesure, a été placée en liquidation judiciaire le 8 novembre 2010. Une quarantaine de magasins ont fermé, laissant un millier de salariés sans emploi. Consulté un mois après, le site www.vogica.fr avait entièrement disparu ; à sa place, figurait un avis de 12 lignes signé du mandataire chargé de la liquidation judiciaire.

Où les anciens salariés de l'entreprise peuvent-ils retrouver la mémoire numérique de leur entreprise et, par exemple, montrer que six mois avant sa fermeture, la rubrique « offres d'emploi » du site proposait encore 28 postes ? Où les chercheurs qui, dans 10 ou 30 ans, écriront l'histoire du déclin du tissu industriel français retrouveront-ils les traces du développement, puis de la chute de cette société ? Où, enfin, les amateurs

de publicités retrouveront-ils les clips qui ont fait la popularité de l'enseigne ?

Toutes ces informations sont conservées à la BnF, dans ses « archives de l'Internet ». L'histoire du site de *Vogica* illustre la problématique fondamentale du patrimoine : ce qui semble anecdotique et sans valeur aujourd'hui peut, un jour ou l'autre, devenir une source précieuse pour l'histoire collective et intéresser nombre de nos concitoyens. Et le Web étant devenu un média de masse, où publier est désormais à la portée de chacun, le patrimoine national ne peut plus se limiter à la production des élites et des spécialistes. Pour la première fois, il est possible de l'étendre aux propos des citoyens les plus ordinaires. Le dépôt légal du Web répondrait-il au défi que représente l'invention d'une mémoire de masse ?

Le dépôt légal, né d'une ordonnance royale de François I^{er} en 1537, a permis de réunir la plus grande part des collections de la BnF. Il oblige tout éditeur, imprimeur, distributeur ou importateur à déposer des exemplaires de chaque document qu'il édite, imprime, produit, distribue ou importe sur le territoire français, à l'un des organismes dépositaires (BnF, INA, etc.). Le dépôt légal s'est étendu à tous les supports et modes de diffusion, jusqu'à la révolution numérique. Ainsi, depuis le vote de la loi de 2006, la BnF archive chaque jour, au moyen de robots moissonneurs, des centaines de sites Web. À chaque connexion, le robot d'archivage s'identifie auprès des serveurs de l'éditeur, déclarant son identité et les raisons légales de sa visite. La BnF archive les contenus récoltés dans de vastes baies de stockage – ses magasins numériques. Des copies des sites de journaux, de revues savantes, de créations artistiques, mais aussi des blogs, des jeux en ligne, des sites publicitaires ou commerciaux côtoient désormais les vénérables collections d'estampes, de cartes et de manuscrits. Et depuis 2008, ces archives sont accessibles au public.

À la différence du dépôt légal usuel, le dépôt légal des sites Web n'implique aucune démarche active de la part des éditeurs ou des webmasters. Les collectes se font automatiquement au moyen d'un robot du nom de *Heritrix*. C'est un logiciel libre, comme

Le circuit de traitement d'une archive Web

Une archive Web se constitue en trois étapes – collecte, conservation et signalement. Les défis à chaque étape sont la volumétrie, l'hétérogénéité et la volatilité des données : l'obtention d'une image fidèle du Web exige un parc de serveurs robustes et capables d'archiver un grand nombre de fichiers très variés dans un temps limité.

Les fichiers sont récoltés au moyen de robots logiciels nommés *crawlers* ou *spiders*, le plus utilisé par les bibliothèques nationales étant *Heritrix*. À partir d'une liste d'adresses URL, ces robots copient tous les contenus qu'ils rencontrent, comme le ferait un internaute surfant sur tous les liens d'un site. Les bases de données, les contenus payants ou accessibles par iden-

tifiant, etc., sont un obstacle, et la moisson se limite souvent à la surface du Web. Le paramétrage des robots et leur surveillance par des professionnels permettent de choisir la fréquence et la profondeur de capture des sites.

Les archives sont stockées dans un format (WARC) qui regroupe les fichiers et les métadonnées associées à la collecte (par exemple pour dater l'archive). La sauvegarde se fait sur disques ou sur bandes, dans des entrepôts numériques bien maintenus (copie sur plusieurs sites par exemple). Pour suivre l'évolution des formats et des logiciels, des campagnes de migration (d'un format de fichier vers un format plus récent) ou d'émulation (par la reconstitu-

tion artificielle de l'environnement initial de publication des sites) sont prévues.

Des techniques d'indexation permettent la recherche et la visualisation dans les archives. L'indexation en plein texte permet de rechercher par mots, comme avec un moteur de recherche en ligne. L'application *Wayback Machine* « remonte le temps » en cherchant, à partir de l'adresse d'un site et d'une frise chronologique cliquable, toutes les anciennes versions du site. D'autres techniques offrent aux chercheurs des traitements plus pointus (fouille de données, de liens, analyses sémantiques de corpus...) et des représentations cartographiques et dynamiques des données.