Quand l'histoire s'écrit sur la toile. Archiver le Web : genèse, outils et enjeux

Anouk Chapuis

1 Introduction

Depuis sa création en 1991, le Web s'est affirmé comme l'un des principaux canaux de communication, soulevant de nouveaux enjeux. La question de sa préservation, notamment, apparaît comme une préoccupation majeure, dont les médias généralistes se sont fait ces dernières années l'écho [1]. La littérature scientifique a très tôt alerté et justifié de la nécessité d'archiver le Web. Outil de publication dynamique, ouvert et participatif, le Web est un « réservoir » à informations à valeur culturelle, historique et scientifique [Kahle (1997)]. Dans un article paru en 2011, la chercheuse H. Hockx-Yu résumait : « It plays an undisputed important role in modern society, fundamentally changing the way we live and communicate. Its impact has been felt in how we publish, learn, teach and research, and many other areas of human activities » [Hockx-Yu (2011)]. Or, si l'information en ligne se crée facilement, elle disparaît avec la même promptitude. Une étude réalisée entre 2004 et 2006 a estimé qu'en l'espace de deux ans, à peu près 50 pourcents des pages Web avaient disparu [Koehler]. Une autre, elle aussi réalisée en 2004, a constaté qu'au cours d'une semaine, 35 à 40 pourcents des pages Web avaient vu leur contenu être modifié [Fetterly *et al.* (2004)]. Bien souvent, une mise à jour écrase la version précédente, sans que celle-ci n'ait été conservée. Le caractère éphémère du Web a d'ailleurs largement contribué à alimenter la peur d'un « dark age » numérique au tournant des années 2000. La nature ambivalente du Web – facilité d'édition et obsolescence des contenus édités – fait ressentir son archivage comme une nécessité d'autant plus urgente. M. Dougherty, dans deux articles publiés en 2010 et en 2014, et menés sur base d'entretiens avec des chercheurs issus des sciences humaines, a montré que le Web archivé rencontre les besoins de nombreuses disciplines (linguistique, sciences politiques, etc.) [Dougherty *et al.* (2010), Dougherty et Meyer (2014)]. En 2003,

1. Le 9 avril 2020, le magazine en ligne Slate titrait par exemple "La bibliothèque nationale américaine archive tous les mèmes en ligne".

le Web archivé a permis la parution d'une étude sur l'évolution des hyperliens [Kraft *et al.* (2003)] ; en 2005, les archives Web du 11 septembre 2001 ont servi de base à une analyse des expressions du deuil et des structures en ligne de l'engagement civique en temps de crise [Foot (2005)]. Les exemples pourraient être multipliés.

Né dans le sillage de l'Internet Archive (1996), initiative pionnière, l'archivage du Web est un champ de recherche jeune et pluriel, mêlant des approches très technicistes à des considérations d'ordre épistémologique, éthique, juridique et philosophique. Le sujet a fait l'objet d'une littérature abondante, qu'il n'est pas possible, dans le cadre de ce travail, de synthétiser de manière exhaustive. Quelques grandes tendances peuvent néanmoins être dégagées. La littérature sur l'archivage du Web a longtemps été dominée par des questions de méthodologie. Les travaux du Danois N. Brügger en particulier ont mis l'accent sur l'archive Web en tant que source de l'historien, et sur ce qui distingue ce matériel d'archives plus traditionnelles [Brügger (2012)]. Ces préoccupations méthodologiques reflètent sans doute le besoin de réfléchir et de légitimer un champ de recherche encore récent. Depuis, le questionnement a été élargi, notamment en France, où des chercheurs comme G. Illien et Fr. Musiani ont donné à l'archivage du Web un ancrage politique, en interrogeant par exemple la répartition géographique des projets d'archivage, ou les enjeux mémoriels qui les sous-tendent [Illien (2011), Musiani et Schafer (2017), Musiani *et al.* (2019)]. Les publications se sont également faites plus vulgarisatrices, traduisant la volonté d'ouvrir les archives Web à un public plus large, à l'image de l'article de J. Niu à destination d'étudiants en sciences de l'information [Niu (2012)]. Ce travail se propose de présenter les principaux acquis de la littérature et s'intéressera en particulier à trois éléments : la chronologie des projets notables d'archivage du Web, quelques-uns des outils les plus couramment utilisés (le crawler Heritrix, le format WARC et l'interface Warclight) [2] et les enjeux que soulève l'archivage du Web, sur le plan technique, sur le plan de l'authenticité et de l'accessibilité et sur le plan juridique.

2. Ces outils ont notamment été retenus dans le cadre du projet belge PROMISE, lancé en 2017.

2 L'archivage du Web : une histoire récente

2.1 Un projet pionnier : l'Internet Archive (1996)

Si un archivage du Web « primitif », essentiellement pratiqué par des particuliers (chercheurs, chefs d'entreprise, familles, etc.) – et orienté vers des besoins très individuels –, a rapidement suivi la création du Web au début des années 1990 [Brügger (2011)], la première tentative d'archivage systématique du Web revient à l'Américain Brewster Kahle. Ingénieur diplômé du MIT (Massachusetts Institute of Technology), il fonde en 1996 l'entreprise d'informatique Alexa Internet et, la même année, l'Internet Archive [Brown (2006)]. Organisation à but non lucratif, l'Internet Archive a pour vocation de préserver la mémoire du Web mondial. À l'origine, le projet naît de la volonté d'archiver les sites Web des candidats à l'élection présidentielle de 1996 mais assez vite, il adopte une stratégie de collecte des contenus Web beaucoup plus large (aussi connue sous le nom de « snapshot strategy » ou stratégie de l'instantané dans la littérature). Brewster Kahle ambitionne de créer une « bibliothèque universelle du Web » et donne d'ailleurs en 2003, symboliquement, une copie des collections de l'Internet Archive, dont le siège est à San Francisco [3], à la Bibliotheca Alexandrina, en Égypte [Brown (2006), Mussou (2012)].

Dix ans après sa fondation, la plateforme avait collecté quelques 50 billions de pages Web, à l'usage de 70 000 visiteurs par jour ; aujourd'hui, elle rassemble plus de 500 billions d'objets [Webster (2017)]. Ces objets sont collectés par un crawler, d'abord celui d'Alexa – qui a été depuis vendue à Amazon –, puis par une version open-source (Heritrix), spécifiquement développée pour l'Internet Archive. Depuis 2001, les collections sont rendues accessibles via la Wayback Machine, interface qui permet de rechercher les contenus par URL. L'accès aux contenus n'est pas restreint, l'Internet Archive ayant choisi (à l'instar de Google Books) de recourir à l'opt-out. Les contenus sont soustraits de la consultation à la seule demande de leurs auteurs et/ou éditeurs. Projet titanesque, l'Internet Archive exerce une influence forte sur les pratiques d'archivage du Web à l'échelle mondiale. Nombreuses sont les initiatives, notamment nationales, à utiliser le robot Heritrix, ainsi que le format de stockage WARC, lui aussi pensé par les ingénieurs de l'Internet Archive, participant ainsi à en faire des standards. Premier fonds d'archives Web – et seules archives Web pour une majorité de domaines nationaux –,

3. Les deux PetaBytes des collections compressées, eux, sont stockés dans la Petabox à Santa Clara, en Californie.

le projet ne doit pas pour autant éclipser les autres programmes d'archivage qui ont fleuri à la même époque.

2.2 Les projets publics à partir du milieu des années 1990

En Europe, les pays nordiques sont parmi les premiers à investir l'archivage du Web, notamment la Suède à travers son projet Kulturarw3, lancé en 1996 et confié à la Bibliothèque nationale. Kulturarw3 a pour objectif la collecte du domaine national suédois. En 2005, il rassemblait 350 000 sites Web [Brown (2006)] et fait, encore aujourd'hui, partie des plus grandes collections au monde. Comme l'Internet Archive, le projet a opté pour une stratégie de collecte large, rendue possible par l'usage du crawler Heritrix. Mais contrairement à l'Internet Archive, et pour la première fois, Kulturarw3 impose au périmètre de collecte des frontières nationales. Autre différence majeure, les collections sont uniquement consultables sur site, en bibliothèque. Un an plus tard, en 1997, naît le Nordic Web Archive, projet collaboratif réunissant les bibliothèques nationales du Danemark, de Finlande, d'Islande, de Norvège et de Suède [Brown (2006)]. Le Nordic Web Archive a pour missions le partage et la coordination des expériences des différentes bibliothèques membres en matière d'archivage du Web, ainsi que le développement d'outils. Il a entre autres contribué à développer le NWA Toolset, logiciel open-source d'accès au matériel Web archivé, dont s'inspire directement le logiciel WERA (Web Archive Access) de l'IIPC (International Internet Preservation Consortium). La même année, en 1997, la Finlande lance EVA, son programme d'archivage du domaine national, suivie en 2002 par le Danemark et son Netarchive.dk [Brown (2006)]. Le pilote est géré conjointement par la Bibliothèque royale et la bibliothèque universitaire d'Aarhus. L'approche danoise est hybride, mêlant à une collecte large une collecte beaucoup plus ciblée. La Bibliothèque royale collecte régulièrement le domaine national, ici encore à l'aide d'Heritrix, tandis que la bibliothèque universitaire d'Aarhus s'occupe de créer des collections thématiques, beaucoup plus sélectives.

En 1996, la Bibliothèque nationale d'Australie avait déjà choisi, dans le cadre de son projet PANDORA, d'appliquer une stratégie de collecte restrictive, se démarquant alors des programmes existants par l'exigence de sa sélection [Brügger (2011), Webster (2017)]. Seul un nombre limité de sites Web, sélectionnés un à un, sont destinés à être archivés. Des experts sont chargés de créer des collections structurées par thème (par exemple, le changement climatique), par événement (par exemple, les élections) ou par struc-

ture (blogs, sites d'informations, etc.). Cette logique sélective a été motivée par un écueil juridique : pour rendre publics les contenus archivés, la Bibliothèque a au préalable besoin des autorisations de leur créateur et/ou éditeur. Celles-ci n'auraient pas pu être obtenues dans le cadre d'une collecte plus large. PANDORA est donc l'un des rares projets nationaux à mettre en ligne ses archives, avec l'accord des propriétaires des documents originaux [Brown (2006)]. Le projet, né concomitamment à l'Internet Archive, a également développé son propre crawler (PANDAS) [Brown (2006)]. Privilégiant la profondeur à la superficie, la stratégie de collecte de PANDORA exige davantage d'effectifs et de moyens financiers, ce qui explique que de nombreuses bibliothèques choisissent de ne recourir à la collecte sélective qu'en complément d'une collecte plus large, à l'image du projet danois Netarchive.dk. Depuis le milieu des années 1990, le périmètre et les pratiques d'archivage se sont diversifiées, un nombre croissant de bibliothèques et d'universités entreprenant leur programme : en 1999 la Nouvelle-Zélande, en 2000 les États-Unis et la République tchèque, en 2001 la Norvège, en 2002 la France et le Japon, etc.

2.3 Une institutionnalisation croissante au tournant du 21e siècle

La première décennie du 21e siècle a vu l'archivage du Web s'institutionnaliser, à travers la création de consortiums et l'adaptation de la législation sur le dépôt légal. En 1997, le Danemark est l'un des premiers pays à modifier ses textes de lois sur le dépôt légal pour y inclure les documents non imprimés [Webster (2017)]. La même année, la Commission européenne crée NEDLIB (Networked European deposit library) dans le but d'aider les bibliothèques nationales à remplir leurs missions quant au dépôt légal, y compris concernant les publications numériques [Brown (2006)]. Fruit de la coopération de neuf bibliothèques, le projet NEDLIB teste d'abord plusieurs crawlers existants avant de développer le sien, utilisé dans le cadre d'un certain nombre de programmes, notamment aux Pays-Bas, en Estonie et en Islande. Il tend toutefois à être remplacé par Heritrix. Les pays sont de plus en plus nombreux à élargir leur dépôt légal aux publications en ligne [Musiani *et al.* (2019)]. C'est le cas de la France en 2006, du Royaume-Uni en 2013, etc. En 2017, la Belgique a lancé un projet pilote d'archivage du Web [Pretoro *et al.* (2020)] qui a abouti, l'année suivante, à l'adoption du principe du dépôt légal numérique [Vandepontseele (2017)].

En juillet 2003, la coopération prend une dimension intercontinentale avec la création de l'International Internet Preservation Consortium [Illien (2011)].

L'Internet Archive et une dizaine de bibliothèques nationales européennes, nord-américaines et australiennes se réunissent en un consortium, dans le souci de garantir l'interopérabilité des collections. Le Web étant un médium transnational, son archivage appelait tôt ou tard une forme de concertation. L'IIPC cherche à créer une communauté de formats, de logiciels et de normes, pour permettre cette interopérabilité, mais cherche également à prévenir la dispersion des efforts et la redondance des contenus archivés. Ses outils, largement empruntés à l'Internet Archive et rassemblés dans un « toolkit » ou « toolset », ont la particularité d'être open-source. L'IIPC compte aujourd'hui une cinquantaine de membres. Ceux-ci sont cooptés et sont tenus de s'acquitter d'une cotisation annuelle (entre 2000 et 8000€ selon leur budget). Ils doivent également pouvoir justifier de réalisations significatives en matière d'archivage du Web. L'IIPC est une organisation quasiment virtuelle, dont les membres ne se rencontrent en présentiel qu'une à deux fois par an ; elle ne dispose ni d'un siège, ni de personnel salarié permanent, son fonctionnement dépendant du volontariat. Elle a été rejointe en 2017 par la Belgique.

3 Les outils

3.1 La collecte des contenus : le cas d'Heritrix

La méthode d'archivage la plus répandue consiste, à l'aide d'un robot (ou « crawler »), à collecter le contenu depuis des serveurs Web distants (« remote harvesting »). Le crawler imite le comportement humain pour interagir avec un serveur : il lui envoie une requête http et stocke le contenu qu'il reçoit en retour [Brown (2006)]. Le comportement du crawler est dicté par une liste d'URLs (« seed list ») à visiter. Le crawler se rend sur la première URL de la liste, collecte la page Web associée, puis identifie tous les hyperliens qu'elle contient et les ajoute à sa liste. La profondeur du crawl et le nombre d'hyperliens à suivre sont paramétrables. Il existe aujourd'hui une grande variété de softwares d'archivage, aussi bien propriétaires qu'en libre accès [Brown (2006)]. Les trois crawlers les plus utilisés sont cependant HTTrack, développé par Xavier Roche et un temps utilisé dans le cadre du projet PANDORA, NEDLIB Harvester, développé par le NEDLIB Project et dont le développement a pris fin en 2002, et Heritrix. Lancé au début de l'année 2003, Heritrix est le crawler développé par l'Internet Archive, en partenariat avec le Nordic Web Archive. Aujourd'hui, le développement d'Heritrix est soutenu par l'IIPC, qui a intégré le robot à son « toolkit ». L'avantage de la collecte à distance est qu'elle requiert une infrastructure relativement légère, c'est-à-dire un software à installer sur un ordinateur, une connexion internet

et un espace de stockage suffisant [Brown (2006)]. Surtout, les possibilités d'application sont très larges, un crawler pouvant collecter des contenus Web aussi bien en très grande qu'en très petite quantité. Enfin, la méthode ne nécessite pas l'autorisation du propriétaire du site Web, ce qui garantit au responsable de la collecte une certaine autonomie.

Selon ses concepteurs, Heritrix a été pensé comme un crawler open-source dans le but d'encourager la coopération entre institutions, avec une architecture modulable et extensible qui facilite la personnalisation et les contributions extérieures [Mohr *et al.* (2004)]. Heritrix est implémenté en Java, langage supporté par Linux, Windows et Macintosh OS X [Mohr *et al.* (2004)]. Le crawler de l'Internet Archive est défini par trois composants essentiels : le champ d'application (« scope »), la frontière (« frontier ») et les chaînes de processeurs (« processor chains ») [Mohr *et al.* (2004)]. Le champ d'application comprend les URLs « d'amorçage » utilisées pour démarrer le crawl. La frontière est un composant qui suit les URLs programmées pour être collectées ; le composant est également responsable de la sélection de la prochaine URL à visiter et empêche la programmation redondante des URLs déjà programmées. Les chaînes de processeurs s'occupent de la collecte en elle-même : ils envoient notamment la requête http et stockent la réponse dans un format de fichier WARC. Malgré que la méthode soit éprouvée et que les résultats impressionnent par leur taille, la collecte à distance a aussi des faiblesses [Brown (2006)]. Les larges volumes de données et la vitesse de collecte eux-mêmes peuvent poser problème : il faut que le reste du processus d'archivage puisse suivre le rythme au risque de voir des retards s'accumuler. L'usage d'un crawler comme Heritrix implique une configuration minutieuse pour obtenir des résultats convaincants ; si les connexions au serveur sont trop rapides, celui-ci risque par exemple de crasher. Surtout, les crawlers sont incapables de collecter certains types de contenus, limite dont il sera question plus loin.

3.2 Le stockage des contenus : le format WARC

La plupart des organismes d'archivage stockent le matériel Web collecté dans le format de fichier WARC (Web ARCHive). WARC est un format dit « conteneur » qui combine plusieurs ressources numériques dans un fichier d'archivage agrégé, auquel sont associées des informations de contexte. Le format a été créé par l'Internet Archive et la California Digital Library. L'IIPC a participé à le faire reconnaître comme norme ISO en 2009, récemment republiée en 2017. Un fichier WARC est lui-même un enchaînement

(ou « concaténation » dans la littérature spécialisée) de plusieurs enregistrements WARC. Chaque enregistrement est composé d'un en-tête (« header ») et d'un bloc de contenu (« payload »). L'en-tête reprend des champs obligatoires tels que la date, le type et la longueur de l'enregistrement ; le bloc de contenu rassemble quant à lui les ressources de n'importe quel format (audio et vidéo inclus) sous forme binaire. La force du format WARC est qu'il peut gérer à la fois les données collectées et celles documentant la collecte, générant des métadonnées relativement riches. Le format exige toutefois une maintenance complexe et lourde, peu facile à mettre en place au sein des plus petits organismes d'archivage. L'accès à l'archive n'est en effet pas direct ; il passe par deux « couches » qui s'y superposent : un système d'indexation des fichiers WARC et un serveur Web qui délivre les archives indexées à l'usager [Masanés (2006)].

3.3 L'accès aux contenus : l'interface Warclight

Warclight est un produit issu de la plateforme open-source Blacklight et du format de fichier WARC [Ruest *et al.* (2019)]. Il permet à l'usager de rechercher des archives Web indexées sur la plateforme de moteur de recherche Apache Solr. Les archives stockées au format WARC sont organisées dans des index conçus pour être compatibles avec Apache Solr, plateforme particulièrement performante, utilisée par exemple par eBay, Netflix ou Disney. Les index sont créés grâce à l'outil Web Archive Discovery de la UK Web Archive, selon une variété de champs interrogeables, comme le titre, le corps, le type de contenu, etc. Mais Apache Solr ne proposant pas d'interface usager directe, il faut recourir à Blacklight, interface conçue pour explorer les documents indexés, ou dans le cas des archives Web, à la nouvelle interface Warclight. Cette interface propose des fonctionnalités de recherche intéressantes, plus élargies que celles de la Wayback Machine de l'Internet Archive. Pour l'instant, l'outil, pensé par un groupe de six universités canadiennes, est encore peu répandu et sera sans doute amené à beaucoup évoluer dans les prochaines années. Il vaudra certainement la peine de suivre les résultats qu'apportera ce nouvel outil dans le cadre des projets qui l'ont tout récemment adopté, comme le projet belge d'archivage du Web démarré en 2017.

4 Les enjeux actuels

4.1 Les limites techniques

Objet complexe et protéiforme, le Web confronte les archivistes à certaines difficultés techniques. Dans un guide destiné aux professionnels de l'information, l'archiviste britannique A. Brown a listé les problèmes les plus courants qu'un projet d'archivage du Web peut rencontrer [Brown (2006)]. Certains d'entre eux sont anodins et des solutions relativement faciles à mettre en œuvre existent. C'est par exemple le cas lorsqu'une page Web archivée s'affiche sous forme de texte. Le problème vient le plus souvent du fait que le crawler a été identifié par le serveur comme étant un navigateur « non supporté », auquel cas il est possible de faire croire au serveur qu'il s'agit bel et bien d'un navigateur pris en charge, à l'aide d'outils d'imitation. Parfois, les pages sont collectées dans un mélange aléatoire de langues ; manipuler la gestion des cookies par le crawler, ou le paramétrer de façon à ce qu'il ne suive pas le lien « langues » sur la page d'accueil, suffit en général à rectifier l'anomalie. Mais dans d'autres cas, plus sérieux, il n'existe pas de réponse simple. Les technologies disponibles se heurtent à des limites qui se traduisent, la plupart du temps, par des archives du Web qui présentent de nombreux « trous ». Ces carences ont plusieurs origines :

— il peut s'agir de liens morts. Le contenu lié est hébergé à l'intérieur d'un domaine qui n'a pas été retenu lors de la sélection des sites Web à collecter ;

— le deep Web reste inaccessible aux crawlers. Celui-ci représentant une part considérable du Web – jusqu'à plus de 90 pourcents selon certaines estimations –, son archivage reste un défi majeur ;

— un robots.txt est présent dans le code source d'une page, excluant certains documents du champ d'action des crawlers ;

— certains contenus sont protégés pour des raisons techniques, juridiques ou de sécurité : une large part du Web est protégée par un mot de passe et reste donc inaccessible aux crawlers. L'essor des réseaux sociaux, où contenus publics, semi-publics et privés se côtoient, accentue la tendance. Le déclin du Web « traditionnel » au profit du Web social accroît le risque que des contenus et protocoles jusqu'alors ouverts soient remplacés par des systèmes fermés et non détectés par les outils d'archivage actuels [Dougherty *et al.* (2010)] ;

— certains objets ne sont pas pris en charge par les crawlers parce que générés dynamiquement (par exemple avec Flash ou JavaScript). Il s'agit surtout de contenus multimédia et de contenus générés sur base d'une interaction avec l'usager.

Pour pallier ces lacunes, des alternatives ont été développées. Il s'agira par exemple d'obtenir les copies des ressources concernées directement auprès du serveur (« direct transfer »), avec l'autorisation de leur propriétaire. Seulement, la démarche, si elle a le mérite d'exister, est très chronophage et donc difficilement applicable dans le cadre d'une stratégie d'archivage qui se veut globale. De manière générale, l'évolution rapide des technologies Web nécessite de s'adapter continuellement. Aujourd'hui, les principaux défis sont posés par l'utilisation croissante de contenus exécutables, par l'apparition de nouveaux modes d'affichage des contenus (HTML5, plateformes mobiles, etc.) et par le développement du Web sémantique. Le corollaire de cette évolution rapide des technologies Web est l'obsolescence très forte qui les frappe, qui concerne aussi bien l'hardware que les softwares, les systèmes d'exploitation et les formats de fichier, et qui peut affecter la préservation à moyen et à long terme des archives. La durée de vie limitée des technologies explique sans doute la place importante que la littérature accorde à la nécessité de penser en amont de tout projet une stratégie de préservation et de mettre en place une politique de contrôle et d'évaluation.

4.2 L'authenticité

Un problème est unanimement évoqué par les chercheurs : celui du décalage qui peut apparaître entre la version originale et la version archivée d'une page Web. Idéalement, l'archive doit être isomorphe à l'original mais dans les faits, c'est rarement le cas. Très souvent, des contenus sont manquants, pour les raisons techniques évoquées plus haut. Les crawlers peinent également à reproduire certaines fonctionnalités. Pendant longtemps, il n'a pas été possible d'effectuer une recherche par mot-clé à l'intérieur d'un site Web archivé par l'Internet Archive. Aujourd'hui, cette fonctionnalité est proposée mais reste limitée à la page d'accueil du site [Musiani *et al.* (2019)]. La façon dont l'archive s'affiche peut également différer par rapport à l'original : il suffit que la police d'une page n'ait pas été inscrite dans son code source mais ait plutôt été utilisée par défaut pour que ce soient les paramètres établis dans le navigateur de l'usager actuel qui déterminent la police de la page Web archivée [Musiani *et al.* (2019)]. Mais surtout, des inconsistances peuvent apparaître. La version archivée d'un site Web peut combiner des éléments issus de deux versions du site – ou plus – qui ont pourtant existé à des moments différents [Brügger (2005)]. Le problème vient du fait que la fréquence des mises à jour du site est supérieure à la rapidité du processus d'archivage, autrement dit que le site continue à être rafraîchi alors que son archivage est en cours. N. Brügger en donne un exemple :

« Pendant les Jeux Olympiques de Sydney de 2000, je voulus sauvegarder le site Web du quotidien danois JyllandsPosten. Je commençai au premier niveau, avec la page d'accueil, sur laquelle on pouvait lire que la joueuse de badminton danoise, Camilla Martin, allait jouer en finale trente minutes plus tard. Mon ordinateur mit environ une heure pour sauvegarder ce premier niveau, laps de temps au bout duquel je décidai de télécharger le second niveau, les "Jeux Olympiques de 2000". Mais sur la page d'accueil de cette section, je pouvais déjà lire les résultats de la finale de badminton (la joueuse danoise avait perdu). L'état du site Web – dans son ensemble – n'était plus le même que quand j'avais commencé. Il avait subi des transformations durant le temps pendant lequel il avait été archivé, et je pouvais maintenant découvrir les résultats du match sur la même page d'accueil où le match avait été annoncé quelques minutes auparavant [Brügger (2012)]. »

L'asymétrie entre la version originale d'un site Web et sa version archivée pose les questions de l'authenticité et de l'intégrité de l'archive. Des instruments ont été développés pour visualiser et identifier les défauts de consistance, par exemple dans le cadre du projet LiWA [Hockx-Yu (2011)]. Mais en l'absence d'une technologie qui permettrait de parer ces défauts, il est nécessaire que les organisations en charge de l'archivage définissent ce qui, dans un contexte numérique, est « authentique » et « intègre », selon leurs besoins et leurs missions. Elles peuvent pour cela s'appuyer sur des textes juridiques (au niveau de l'Union européenne, avec le règlement eIDAS ; au niveau belge, avec la loi sur le numérique) [Pretoro *et al.* (2020)] et sur un champ de recherche relativement récent : la diplomatique numérique, qui fait la part belle aux questions de gestion des métadonnées et de documentation. A. Brown recommande également de signaler aux usagers toute divergence entre la version originale et la version archivée, ceux-ci n'ayant pas toujours conscience des écarts qui peuvent les séparer [Brown (2006)].

4.3 L'accessibilité

Le manque d'accessibilité des archives du Web est un obstacle qui n'a pas encore trouvé de réponse définitive et qui se manifeste à deux niveaux, celui, individuel, de l'usager, et un autre plus géopolitique. Au niveau de l'usager, la plupart des interfaces n'offrent pas la possibilité d'effectuer une recherche plein texte et sont peu user-friendly (faible intuitivité) [4]. Cela tient au fait que les crawlers développés par l'Internet Archive ont été conçus à un moment

4. Il faut néanmoins signaler que le développement de nouvelles interfaces, comme Warclight, tendent à faciliter et à enrichir de plus en plus les modalités de recherche.

où le Web était un médium nouveau, dont la principale originalité résidait dans son hypertextualité [Ben-David et Huurdeman (2014)]. Les premiers navigateurs permettaient d'accéder au Web à partir d'une URL ou d'un répertoire et la navigation se faisait en suivant les différents liens. Les premiers moteurs de recherche ont par la suite organisé le Web en l'indexant et en construisant eux aussi des répertoires. Les outils de l'Internet Archive sont imprégnés de cette culture de recherche du Web primitif. Mais l'accès aux archives du Web par URL, tel que le proposent la majorité des plateformes, est contraignant car il oblige l'usager à connaître le localisateur exact de la page qu'il souhaite consulter. A. Ben-David et H. Huurdeman, auteurs d'une étude sur les interfaces de recherche des archives du Web, ont très bien formulé le problème : « Put differently, most web archives are not searchable. To consult the web archive through the Wayback Machine, a researcher must know the URL she would like to retrieve from the archive [...]. Thus without the ability to search for keywords or other contextual elements, it would be dificult for future historians to trace the relevant URL as the starting point to studying a theme, an issue or an event using web archives [Ben-David et Huurdeman (2014)]. »

Pour davantage rencontrer les attentes de l'usager, M. Dougherty, autrice de deux enquêtes sur l'utilisation des archives du Web par les chercheurs [Dougherty *et al.* (2010), Dougherty et Meyer (2014)], souligne l'importance d'identifier son profil et de sonder ses besoins. Elle met également en évidence l'importance de faciliter la navigation de l'usager en lui procurant toutes les métadonnées sur le contexte dans lequel le matériel Web a été archivé (date de l'archivage, software utilisé, organisation responsable, etc.). Au début des années 2010, des outils ont été développés par des archives Web nationales, notamment au Japon, au Royaume-Uni, au Portugal et aux Pays-Bas, qui permettent une recherche plein texte [Ben-David et Huurdeman (2014)]. Prometteurs, leur évolution et les possibilités qu'ils laissent entrevoir, requerront d'être suivies. À un niveau plus géopolitique, l'accès aux archives du Web peut être compromis lorsque des gouvernements décident de l'interrompre, de manière plus ou moins prolongée, comme cela a été le cas en Chine en 2014, en Russie en juin 2015 et en Jordanie en 2017 [Musiani *et al.* (2019)] ou lorsque l'infrastructure est insuffisante pour supporter un service d'archivage du Web. La géographie des programmes d'archivage du Web révèle en effet une fracture Nord-Sud marquée. S'il est important de rappeler le contexte international dans lequel évolue l'archivage du Web, en faire une analyse détaillée sortirait cependant du cadre que ce travail s'est fixé.

4.4 Les questions juridiques

L'archivage du Web soulève des problématiques juridiques aussi variées que celles touchant au droit à la propriété intellectuelle, au droit à la protection des données et à la vie privée et à la légalité des contenus collectés [Brown (2006)]. La plupart des législations nationales protègent les contenus Web, que ce soit au travers du copyright anglo-saxon ou au travers des droits d'auteur francophones. Si les législations diffèrent d'un pays à l'autre, la plupart des auteurs pointent le caractère obsolète de ces législations qui continuent à fonctionner sur la base d'une distinction entre l'œuvre originale et sa copie. Or, dans un contexte numérique, cette distinction est peu opérante : puisque le processus d'archivage fait subir au contenu archivé une transformation, faut-il considérer l'archive comme une copie ou comme une nouvelle version, une adaptation de l'original ? C'est pour cette raison que la plupart des projets d'archivage du Web, y compris le projet PROMISE, comportent un volet juridique non négligeable. La problématique du droit à la protection des données personnelles, du droit à l'oubli et du droit de rectification doit également être prise en compte, même si une partie importante du Web est protégée par un mot de passe. Des contenus illégaux (diffamation, obscénité, promotion d'activités illégales, etc.) peuvent également être collectés lors de l'archivage et nécessitent de penser des solutions (contrôle qualité, accès limité, politique de retrait, avertissement, etc.).

5 Conclusion

Un survol rapide de l'histoire de l'archivage du Web montre des pratiques diversifiées et, souvent, hybrides. Le projet danois Netarchive.dk allie par exemple à une collecte des contenus large une collecte beaucoup plus sélective. Chaque année, quatre « instantanés » du Web danois sont collectés par la Bibliothèque royale. Les instantanés sont complétés par la collecte quotidienne de 80 sites web, soigneusement sélectionnés par les équipes du projet, et par la collecte de sites Web relatifs à deux ou trois événements annuels, en fonction de l'actualité. De même, les acteurs de l'archivage du Web se sont, au fil du temps, diversifiés. L'apparition de services d'externalisation dans les premières années du 21e siècle a entraîné la démocratisation de l'archivage du Web, le faisant entrer dans les universités, les écoles, les églises, les organisations commerciales, etc. L'European Web Archive, organisation sans but lucratif fondée en 2004 à Amsterdam, en coopération avec l'Internet Archive, propose de tels services. Aux États-Unis, l'Internet Archive a développé en 2006 son propre service, Archive-It, accessible via une application

Web qui permet aux usagers de gérer facilement leurs archives. Des initiatives commerciales ont suivi (Hanzo Archives, Pagefreezer, Aleph Archives, etc.).

À côté des grands acteurs de l'archivage que sont l'Internet Archive et les bibliothèques nationales, de petits groupes d'activistes peuvent également prendre l'initiative d'un programme d'archivage, dans le but de documenter leur action politique et/ou sociale. La plupart des navigateurs proposent maintenant la possibilité d'enregistrer les textes et les images d'une page Web, rendant l'archivage accessible aux particuliers, à très petite échelle. S'il existe des échanges entre tous ces acteurs – comme l'adoption généralisée du crawler Heritrix en atteste –, il y a également une forte segmentation des archives Web. Il n'existe pas nécessairement de passerelles entre les différentes collections. Cela tient à la multiplicité des cadres juridiques, institutionnels et des modalités d'accès. C'est pourquoi la coopération et l'open source auront sans doute, à l'avenir, un rôle important à jouer dans le développement d'outils à la fois performants et respectueux des critères d'interopérabilité. L'archivage du Web étant un champ encore jeune, dont les outils ne sont pas encore tous arrivés à maturité, son évolution restera à suivre dans les prochaines années.

Références

Anat Ben-David et Hugo Huurdeman : Web Archive Search as Research : Methodological and Theoretical Implications. *Alexandria*, 25(1-2):93–111, août 2014. ISSN 0955-7490. URL `https://doi.org/10.7227/ALX.0022`. Publisher : SAGE Publications Ltd.

Adrian Brown : *Archiving Websites : A Practical Guide for Information Management Professionals*. Facet Publishing, juin 2006. ISBN 978-1-85604-553-7. Google-Books-ID : 7NYqDgAAQBAJ.

Niels Brügger : *Archiving websites : general considerations and strategies*. Center for Internetforskning, Aaarhus, 2005. ISBN 978-87-990507-0-3.

Niels Brügger : Web Archiving – between Past, Present, and Future. pages 24–42. avril 2011. ISBN 978-1-4443-1486-1.

Niels Brügger : L'historiographie de sites Web : quelques enjeux fondamentaux. *Le Temps des medias*, n° 18(1):159–169, juin 2012. ISSN 1764-2507. URL `https://www.cairn.info/revue-le-temps-des-medias-2012-1-`

page-159.htm. Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : Nouveau Monde éditions.

Meghan Dougherty et Eric T. Meyer : Community, tools, and practices in web archiving : The state-of-the-art in relation to social science and humanities research needs. *Journal of the Association for Information Science and Technology*, 65 (11):2195–2209, 2014. ISSN 2330-1643. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23099. _eprint : https ://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23099.

Meghan Dougherty, Eric T. Meyer, Christine McCarthy Madsen, Charles van den Heuvel, Arthur Thomas et Sally Wyatt : Researcher Engagement with Web Archives : State of the Art. SSRN Scholarly Paper ID 1714997, Social Science Research Network, Rochester, NY, août 2010. URL https://papers.ssrn.com/abstract=1714997.

Dennis Fetterly, Mark Manasse, Marc Najork et Janet L. Wiener : A large-scale study of the evolution of Web pages. *Software : Practice and Experience*, 34(2):213–237, 2004. ISSN 1097-024X. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.577. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.577.

Kirsten Foot : Web-based memorializing after September 11 : Toward a conceptual framework. *Journal of Computer-Mediated Communication*, 11(1):72, 2005. ISSN 10836101.

Helen Hockx-Yu : The past issue of the web. *In Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 1–8, New York, NY, USA, juin 2011. Association for Computing Machinery. ISBN 978-1-4503-0855-7. URL https://doi.org/10.1145/2527031.2527050.

Gildas Illien : Une histoire politique de l'archivage du web, janvier 2011. URL https://bbf.enssib.fr/consulter/bbf-2011-02-0060-012.

Brewster Kahle : PRESERVING THE INTERNET. *Scientific American*, 276(3):82–83, 1997. ISSN 0036-8733. URL https://www.jstor.org/stable/24993660. Publisher : Scientific American, a division of Nature America, Inc.

Wallace Koehler : A longitudinal study of Web pages continued : a consideration of document persistence. URL http://informationr.net/ir/9-2/paper174.html. ISSN : 1368-1613 Publisher : Professor T.D. Wilson.

Reiner Kraft, Hastor, Enes et Raymie Stata : TimeLinks : Exploring the link structure of the evolving Web. mai 2003.

Julien Masanés : Web Archiving : Issues and Methods. *In* Julien Masanés, éditeur : *Web Archiving*, pages 1–53. Springer, Berlin, Heidelberg, 2006. ISBN 978-3-540-46332-0. URL https://doi.org/10.1007/978-3-540-46332-0_1.

Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery et Michele Kimpton : An Introduction to Heritrix An open source archival quality web crawler. *In In IWAW'04, 4th International Web Archiving Workshop*. Springer Press, 2004.

Francesca Musiani, Camille Paloque-Berges, Valérie Schafer et Benjamin Thierry : *Qu'est-ce qu'une archive du Web ?* OpenEdition Press, février 2019. ISBN 979-10-365-0367-2 979-10-365-0368-9. URL https://library.oapen.org/handle/20.500.12657/25749. Accepted : 2019-06-19 03 :00 :37.

Francesca Musiani et Valérie Schafer : Les archives du Web : gouvernance et identités. *Gazette des archives*, 245(1):203–215, 2017. URL https://www.persee.fr/doc/gazar_0016-5522_2017_num_245_1_5527. Publisher : Persée - Portail des revues scientifiques en SHS.

Claude Mussou : Et le Web devint archive : enjeux et défis. *Le Temps des medias*, n° 19(2):259–266, novembre 2012. ISSN 1764-2507. URL https://www.cairn.info/revue-le-temps-des-medias-2012-2-page-259.htm. Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : Nouveau Monde éditions.

Jinfang Niu : An Overview of Web Archiving. *D-Lib Magazine*, 18(3/4), mars 2012. URL https://digitalcommons.usf.edu/si_facpub/308.

Emmanuel Di Pretoro, Friedel Geeraert, Peter Merchant et Alejandra Michel : PROMISE : Preserving Online Multiple Information : towards a Belgian strategy : final report. 2020. URL https://researchportal.unamur.be/en/publications/promise-preserving-online-multiple-information-towards-a-belgian-. Publisher : Belgian Science Policy Office.

Nick Ruest, Ian Milligan et Jimmy Lin : Warclight : A Rails Engine for Web Archive Discovery. *In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 442–443, juin 2019.

Sophie Vandepontseele : L'organisation par la Bibliothèque royale de Belgique d'un dépôt légal pour les publications numériques belges. *In Monte Artium*, 10:205–216, janvier 2017. ISSN 2031-3098. URL `https://www.brepolsonline.net/doi/abs/10.1484/J.IMA.5.114688`. Publisher : Brepols Publishers.

Peter Webster : Users, technologies, organisations : Towards a cultural history of world web archiving. *Web 25. Histories from 25 Years of the World Wide Web*, pages 179–190, 2017. URL `https://hcommons.org/deposits/item/hc:26187/`. ISBN : 9781433132698 Publisher : Peter Lang.